7-12-2018

Oxford Nanopore Technology: A Promising Long-Read Sequencing Platform To Study Exon Connectivity and Characterize Isoforms of Complex Genes

Gopinath Rajadinakaran
University of Connecticut - Storrs, gopinath.rajadinakaran@uconn.edu

Oxford Nanopore Technology: A Promising Long-Read Sequencing Platform To Study Exon Connectivity and Characterize Isoforms of Complex Genes

Gopinath Rajadinakaran, Ph.D.

University of Connecticut, 2018

The central dogma states that the genetic information contained in DNA flows to RNA through the process of transcription which, in turn, can result in protein synthesis through translation. Alternative splicing is a mechanism by which multiple mRNA isoforms are generated from a single gene. Ultracomplex genes, characterized by their ability to encode hundreds to thousands of isoforms, arise from a combination of multiple splicing events. Our understanding of alternative splicing improved vastly in the past decade due to the advent of next-generation sequencing (NGS) technologies. The NGS technologies are powerful and have enabled scientists to measure the expression of genes and isoforms digitally, assemble genomes, reconstruct transcriptomes and clinicians to cater treatments that are specific to an individual's genetic makeup. While NGS technologies have many strengths, the shorter read lengths generated from these platforms limit their ability to study exon connectivity over long distances and this information is often *inferred* through statistical means rather than direct measurement. Additionally, the repetitive regions in the genome represents a special case where the short reads have inherent difficulty in joining two adjacent different contigs into a scaffold. The third-generation sequencing technologies, characterized by their ability to generate ultra-long reads can be used to address these limitations.

Here, I have used the Oxford Nanopore (ONT) MinION device to first demonstrate the utility of nanopore technology to sequence long reads to identify exon connectivity using the *Drosophila Rdl*, *MRP*, *Mhc* and *Dscam1* genes. I extended this approach to sequence full-length cDNAs generated from SIRV spike-in RNA to determine the quantitative ability of the platform. These experiments demonstrate the ability of ONT platform to deconvolute isoforms and by

sequencing *Drosophila* ultracomplex genes, I also show that ONT can identify previously unannotated exons and RNA editing sites over long distances. By using direct RNA sequencing, I demonstrate the ability to sequence full-length *Eno2* RNA molecules and that a majority of the reads were sequenced full-length.

.

Oxford Nanopore Technology: A Promising Long-Read Sequencing Platform To Study Exon Connectivity and Characterize Isoforms of Complex Genes

Gopinath Rajadinakaran

B.Tech., Anna University, 2010

M.S., Western Kentucky University, 2012

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy at the

University of Connecticut

2018

APPROVAL PAGE

Doctor of Philosophy Dissertation

Oxford Nanopore Technology: A Promising Long-Read Sequencing Platform To Study Exon Connectivity and Characterize Splice Isoforms of Complex Genes

Presented by Gopinath Rajadinakaran, B.Tech., M.S.

Major Advisor

_____

Brenton R. Graveley

Associate Advisor

_____

Gordon G. Carmichael

Associate Advisor

_____

Blanka Rogina

Associate Advisor

_____

Jeffrey Chuang

Associate Advisor

_____

Zhengqing Ouyang

University of Connecticut

2018

ACKNOWLDEGEMENTS

There are so many wonderful people I have met in my academic life each of whom have contributed towards my accomplishments and I would like to express my gratitude to all of them. The Doctor of Philosophy is one of the highest academic degrees one could earn and is not for the faint of the heart. I have survived this rigorous training mainly because of the unwavering support I received from my thesis advisor Brent Graveley. Since the day I joined his lab in 2014, his unique style of mentoring students helped me explore science independently and grow myself as a scientist. In addition, Brent also helped me pursue my dream career path of combining science with business and I am very thankful for Brent's constant support throughout the completion of this program.

Dr. Kream has been instrumental in navigating the intricacies of the PhD/MBA dual degree program and without her help, I would have been stuck wondering what to do next! I want to express my deepest gratitude to the University of Connecticut and UConn Health for providing me with the financial support without which completion of neither of my programs would have been possible.

My committee members Gordon Carmichael, Blanka Rogina, Zhengqing Ouyang and Jeff Chuang have been a great resource and helped me with their constructive feedback throughout my doctoral training - many thanks to each one of you! I want to extend my special thanks to Michelle Cote at the Center for Entrepreneurship & Innovation, UCONN because as much as Brent helped me learn science, Michelle helped me develop entrepreneurial skills.

Every year I look forward to the month of March, because that is when Sara brings us Girl's Scout cookies and on every social occasion, Lijun amazes each one of us every time with her photographic skills. Many thanks to both of you for making this lab feel like a home! Thanks

to Xintao Wei and Mike Duff for helping me build a small digital home on the HPC cluster and I felt like I was a PRO with Xintao's submitJob commands! Thanks to Alex Plocik and Mohan Bolisetty for helping me learn Bioinformatics and all other past and current lab members for making this lab environment friendly and intellectually stimulating. I am very thankful to all the friends I made here at UConn Health for making this academic journey a memorable one!

My journey to this point has been incredible and I owe much to my family, especially, my mom, dad, uncle, wife and brother, for making this a great one. The encouragements I received from each one of you during tough times has kept me going and helped find the strength in me. A special thanks to all my family because, I would not be here where I am now without you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Oxford Nanopore Technology: A Promising Long-Read Sequencing Platform To Study Exon Connectivity and Characterize Isoforms of Complex Genes

History and Evolution of sequencing Technology

In the field of genetics and genomics, the process of sequencing refers to the identification of linear structure at which the two biological macromolecules, namely, deoxyribonucleic acids (DNA) and ribonucleic acids (RNA), that codes information for the synthesis of protein, occur in one dimensional space. Of these three macromolecules, proteins were the first to be sequenced and Sanger deduced the amino acid sequences of insulin chains in early 1950s [Sanger, 1988]. Following this feat in history was the identification of ribonucleic acid sequences from the yeast alanyl-tRNA in 1965 which took approximately three years to sequence 76 bases using 1 gram of the starting material [Holley et al., 1965]. The number of amino acids that make up proteins is larger than the four nucleotides that comprise DNA and RNA and the difference in chemical properties between various amino acids rendered them relatively easy in the sequencing process. Discovery of sequence specific ribonucleases made the sequencing of RNA molecules possible and while methods were developed around 1970 to sequence DNA molecules, prior efforts failed mainly due to the inability to distinguish the four DNA bases based on distinct chemical properties and also because sequence specific nucleases were not discovered for DNA [Hutchinson, 2007]. The first reports for DNA sequencing came in 1968 that described the sequence of cohesive ends in lambda bacteriophage using primer extension methods [Wu & Kaiser, 1968]. Later, Sanger and Coulson published their work on DNA sequencing using the Plus and Minus method in 1975 and this work was followed by the dideoxy termination and Maxam and Gilbert's method in early 1977 that could sequence hundreds of DNA bases at a time [Shendure et al., 2017]. Now, fast-forwarding to 2018, the current modern-day sequencing has revolutionized the field of genomics

with the development of various different technologies over first, second and third generations that led to rapid improvements in the number of bases and length of reads sequenced along with improved sequence quality that together brought the cost of sequencing down more than would have been predicted by Moore's Law [Wetterstrand, 2018; van Dijk et al., 2014]. Below, I will describe in detail the different technologies that were available in the past, discuss the current and future technologies and the related bioinformatic tools used to analyze the sequence data and finally, conclude with sequencing applications.

First Generation Sequencing:

The Plus and Minus Method

In 1975, Sanger and Coulson published their work on sequencing DNA fragments using the Plus and Minus method [Sanger & Coulson, 1975] which was adopted from two other previous works described by Wu & Taylor [1971] on the Minus method to sequence the cohesive ends of lambda phage and by Englund [1972] on the Plus method developed to study T7 bacteriophage. Sequencing DNA molecules using this method involved two steps and the first of which involved labeling the 5'-end of primers using radiolabeled ^{32}P nucleotides. Under asynchronous and slow polymerization of primer, DNA products of varied length can be obtained that can serve as inputs to the next step. The second step involved extending previously labeled primers followed by fractionating the extended products on acrylamide gel up to one nucleotide resolution and visualization by autoradiography.

The second step involves the Plus and Minus reactions and the labeled products being split into a total of eight different reactions. In the Minus reaction, the template and primer are setup in four individual reactions where each reaction excludes one of the four nucleotides. This specific exclusion causes DNA polymerase I to stall at positions where the corresponding nucleotide is missing. In the Plus reaction, four reactions are setup similar to the Minus reaction but each reaction contains only one of the four nucleotides. Upon treating this reaction mixture

with T4 DNA polymerase, the 3’ to 5’ exonuclease activity starts degrading the template and hydrolyzes few nucleotides until an equilibrium is reached between the hydrolysis and synthesis at 3’-end of the DNA. The equilibrium thus achieved between synthesis and hydrolysis prevents the enzyme from hydrolyzing further and it mimics as though the enzyme is stalled at the respective position [Englund, 1972]. The Plus and Minus reactions are treated with restriction enzymes that were originally used to obtain the oligonucleotide primer and fractionating these products on acrylamide gel enable identification of extended fragments at high-resolution that resolved upto one nucleotide difference. For example, in the case of Adenosine Triphosphate (dATP), the extension products from Minus reaction stalls right before adding ATP due to the specific exclusion of A nucleotides in the reaction whereas in the case of Plus reaction, the hydrolysis stops right at the A site because of its specific inclusion. As a result, the bands observed in the Plus reactions are one nucleotide larger than the Minus reactions which enable sequence identification corresponding to each position.

This method is relatively straight forward and simple to use and approximately but shortcomings were also present. The major limitation of this method was its inability to accurately identify the total number of bases where homopolymer sequences are present. In such cases, Sanger and Coulson relied on the distance between bands from the corresponding Plus and Minus reactions and inferred the repeat sequences were longer proportional to the distance. Other limitations include the requirement for high resolution fractionation of primer extended products to reliably identify bases that are one nucleotide apart. In some cases, longer sequences migrated faster than their smaller counterparts and this migration pattern was observed to be worse under non-denaturing conditions and hints at the presence of potential secondary structures. The presence of artifactual bands also caused problems to reliably identify bases but the artifacts that were inconsistent and not reproducible were usually addressable by repeating the sequencing experiments. Finally, this approach requires the use of both Plus and Minus reactions together and neither method can be used alone to reliably identify complete target DNA sequences.

Despite the drawbacks, this method was used to determine the genome sequence of bacteriophage PhiX174 [Sanger et al., 1977] but the sequence determined by the Plus and Minus was further confirmed by chain termination method [Sanger et al., 1978] and Maxam-Gilbert method [1977].

Ribosubstitution Method

The ribosubstitution method utilizes the property of DNA polymerase I in the presence of Mn^{2+} to incorporate ribonucleotides during DNA synthesis [Berg et al., 1963; van de Sande et al., 1972]. After incorporating ribonucleotides into the template strand, either alkali or ribonuclease can be used to degrade DNA strands into smaller fragments at ribo-substituted position and further analysis of the sequences can be performed by fractionation methods similar to Plus and Minus sequencing [Sanger et al., 1973; Barnes 1978; Brown 1978].

The ribosubstitution method involves annealing of a restriction fragment primer to single stranded template DNA and extending this primer with one ribonucleotide and a limited number of labelled deoxynucleotides [Barnes, 1978]. Following this labeling reaction, four reactions are set up where each reaction contains only one of the four ribonucleotides and all four deoxynucleotides and the primer was extended in the presence of Mn^{2+} ions. The ratio of ribonucleotide in each reaction to that of its corresponding deoxynucleotide is adjusted such that the ribosubstitution occurs about 2% at each position for the corresponding base. Using either a chemical or an enzymatic method, partially substituted ribonucleotides are cleaved and the resulting bands of variable sizes are resolved on high-resolution acrylamide gel. The ribosubstitution approach is similar in principle to Maxam and Gilbert method but consumes less time and involves no hazardous chemicals. This method also offers some advantages over the Plus and Minus method in terms of eliminating artifact bands and producing distinct bands for each nucleotide from repeat sequences. However, this method suffers from the major drawback that infrequent ribosubstitution at specific sequences can lead to weak bands in the visualization

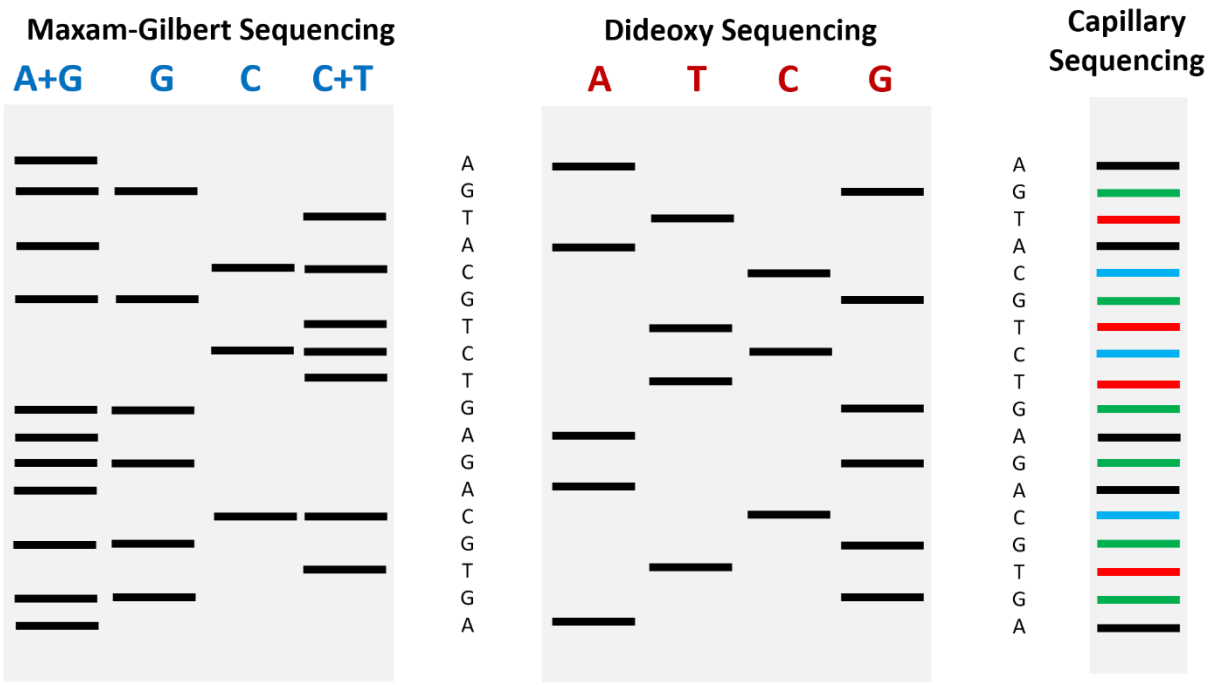
step or these bands can be completely missed out in the gel that results in gaps in the final sequence.

Another variant of the ribosubstitution method developed by Brown [1978] is principally similar to the partial ribosubstitution method except the substitution occurs only at one site immediately following the primer before labeling extension. Following ribosubstitution, the reaction mixtures can be sequenced using the Plus and Minus method and the main advantage of single ribosubstitution over the Plus and Minus method is that the extended products can be obtained by simply treating with either alkali or ribonuclease rather than digesting with restriction enzymes which were found to be inhibited by uncopied single stranded regions in the template DNA. In addition, alkali treatment can generate one fragment as opposed to multiple fragments that are generated when the extended products contain more than one restriction site.

Maxam-Gilbert Sequencing

In early 1977, Maxam and Gilbert developed a chemical method to sequence DNA that selectively targets each of the four nucleotides [Maxam & Gilbert, 1977]. In this approach, DNA is first labeled either at the 5' or 3' end, strands are separated and then treated with either dimethyl sulfate to cleave purines or hydrazine and piperidine to cleave pyrimidines. When the target DNA is treated with dimethyl sulfate, the glycosidic bond formed with guanine and adenine is unstable and can be easily cleaved off following heat treatment. Once the nitrogenous base is removed, the sugar moiety is removed by heat treatment with alkali. This purine specific cleavage using dimethyl sulfate attacks both guanine and adenine but the methylation of guanine is 5-fold faster than with adenine thus resulting in darker bands for guanine and lighter bands for adenines. The glycosidic bond formed with adenine with dimethyl sulfate is weaker than glycosidic bond with guanine and thus treating this reaction mixture with dilute acids preferentially cleaves adenine better than guanine, thus resulting in darker bands for adenine. Thus, these two methods used to cleave purine bases provide complementary information.

Figure 1.1 First generation sequencing technologies. The band pattern observed under for a DNA strand sequenced using Maxam-Gilbert method (A), Sanger's chain-termination method (B) and Capillary sequencing method (C) is shown.


Maxam-Gilbert Sequencing
A+G
G
C
C+T
A
G
T
A
C
G
T
C
T
G
A
G
A
C
G
T
G
A
Dideoxy Sequencing
A
T
C
G
Capillary Sequencing
A
G
T
A
C
G
T
C
T
G
A
G
A
C
G
T
G
A

Hydrazine and piperidine were used to specifically cleave pyrimidine bases. Treating target DNA with hydrazine followed by piperidine results in cleavage products from cytosine and thymine with similar band intensity. In the presence of sodium chloride salt, cleavage at thymine is suppressed and only cytosines are cleaved. Thus, the chemical treatment of DNA chains under four different conditions followed by high-resolution fractionation on polyacrylamide gel enable identification of DNA bases corresponding to each position and this method is limited only by the resolving power of the acrylamide gel.

Sanger Sequencing

Following Maxam-Gilbert's chemical method of sequencing, Sanger and colleagues [1977] published their work in the same year and this method used dideoxy nucleotides that when incorporated into the primer, terminated extension of DNA chains [Atkinson et al., 1969]. Both arabino-nucleosides and dideoxy nucleotides have been used in chain termination experiments but the former method was not suitable for sequencing applications as some mammalian DNA polymerases possess the ability to extend 3'-arabino-nucleosides [Hunter & Francke, 1975]. The dideoxy method is relatively simple and with its ease of use, many shortcomings of the Plus and Minus method were addressed. In addition, the need for strand separation required in Maxam-Gilbert's method was also eliminated with the use of dideoxy nucleotides. The commercial availability of dideoxy nucleotides made this a breakthrough method for DNA sequencing applications and later became widely adopted. Improvements in non-radioactive nucleotides geared towards developing fluorescent dyes coupled with further development of automated sequencing technology allowed ultra-sensitive detection of all four nucleotides in a single reaction tube [Smith et al., 1985; Chidgeavadze & Beabealashvilli 1984; Ansorge et al., 1986]. This single-lane approach allowed the detection of all four bases and completely eliminated the distortions caused between lanes in the four-lane approach where only one fluorophore is used per lane.

The Era of Automated Sequencing and the Human Genome Project

Due to the advantages inherent to the chain termination method such as elimination of radioactive nucleotides and less time consumption, Sanger sequencing was widely adopted to be the method of choice and further improvements such as automation were made for this sequencing method. The development of the Polymerase Chain Reaction (PCR) by Kary Mullis in 1980 and the availability of thermal cyclers and commercial PCR enzymes in 1987 further contributed to the developments in field of sequencing. In 1986, Applied Biosystems first released its automated sequencer ABI370A followed by the 373 and 377 models that increased the sequencing throughput over four-fold. The introduction of capillary technology in 1995 eliminated the need for manual sample loading and a few years later, high-throughput 96-capillary sequencing was introduced. The length of sequenced reads gradually increased from 0.5 kb to about 1 kb with current automated sanger sequencing instruments [Springer, 2006; Hunkapiller et al., 1991; Liu et al., 2012].

Alongside the improvements of automated sequencing, the human genome project (HGP) was started in 1990 with the objective of providing high-quality sequence information about the euchromatic regions present in the human genome. After 13 years of concerted effort between 20 research centers across six countries, the first draft sequence was published in 2001 and a second draft was released two years later in 2003. This project shed light on the 2.85 billion bases present in the human genome (build 35) and identified less than 30,000 protein coding genes [International Human Genome Sequence Consortium, 2001; 2004]. The HGP was expected to take 15 years and projected to cost $3 billion but was completed in 13 years and costed $2.7 billion [NIH, 2018]. Following this success, the National Human Genome Research Institute's (NHGRI) efforts to make the genome sequencing affordable by bringing costs down to a $1000 per genome led to the development of a flurry of technologies, collectively called second-generation sequencing technologies, that dramatically changed the way sequencing was done [Hayden, 2014].

Second Generation Sequencing

The major goals of second or next generation sequencing (NGS) technologies were to offer high through put data at low cost. The NGS technologies can be broadly categorized into two types based on how the sequencing is performed: sequencing-by-synthesis (SBS) and sequencing-by-ligation (SBL). The SBS method involves the identification of bases as the DNA polymerase incorporates individual nucleotides at each position whereas the SBL involves the identification of bases through dinucleotides whereby each position is read twice through ligation process. Each of the NGS technologies discussed below (Table 1.1) have unique advantages and disadvantages that is specific to each platform and it varies across the dimensions of library preparation methods, length of reads sequenced, run times and data throughput [Buermans & den Dunnen, 2014].

454 Platform

The first NGS technology that was commercialized was the 454-pyrosequencing method that is based on the principle of sequencing-by-synthesis (SBS). The 454 technology involves massively parallel sequencing to be performed within small reaction volume of approximately 75 picolitre sized fiber-optic wells on a bead surface [Margulies et al., 2005]. The preparation of the DNA template for sequencing involves attaching the adapter sequences to the fragmented DNA. The adapter sequences contain complementary region to those present on the solid bead surface that facilitates binding of these molecules. Using emulsion PCR, the beads are encapsulated and compartmentalized into tiny droplets such that only one fragment per bead is bound to permit clonal amplification to generate up to 10 million copies [Rothberg & Leamon, 2008]. Following this step, the droplets are collapsed and the beads are deposited into individual wells by washing them over the fiber optic surface containing approximately 1.6 million wells. The bead bound polymerase enzyme is allowed to deposit uniformly across the plate and the sequencing is performed in the next step by supplying one nucleotide at a time, for example, T followed by A, C

Table 1.1 List of different generation of sequencing technologies

Technology	Principle of sequencing	Platforms*	Run time*	Read length*	Throughput*
Sanger sequencing (First)	Sequencing-By-Inhibition	SeqStudio 3500 Series 3730 Series	30 mins 30 mins 20 mins	800 bp 850 bp 900 bp	67 K/day 138 K – 403 K/day 1.38 M - 2.76 M/day
454 (Second)	Sequencing-By-Synthesis	GS Junior GS FLX+	10 hours ~1 day	~700 bp ~1 kb	35 Mb/run 0.7 Gb/run
Illumina/Solexa (Second)	Sequencing-By-Synthesis	iSeq 100 MiniSeq MiSeq NextSeq HiSeq series HiSeq X Nova Seq 6000	9 – 17.5 hours 4 - 24 hours 4 – 55 hours 12 – 30 hours 1 – 6 days ~3 days 16 – 44 hours	2X150 bp 2X150 bp 2X300 bp 2X150 bp 2X150 bp 2X150 bp 2X150 bp	1.2 Gb 7.5 Gb 15 Gb 120 Gb 105 Gb - 1.5 Tb/run 1.6 - 1.8 Tb/run 167 Gb - 6 Tb
SOLiD (Second)	Sequencing-By-Ligation	5500 5500xl 5500 W 5500xl W	6 days 6 days 10 days 10 days	2X60 bp 2X60 bp 2X50 bp 2X50 bp	48 Gb/run 95 Gb/run 120 Gb/run 240 Gb/run
Ion Torrent (Second)		PGM 300 series Ion PI chip Ion 500 series	2 - 7 hours 2 – 4 hours 3 – 22 hours	Up to 400 bp 200 bp 200 – 600 bp	100 Mb – 2 Gb/run Up to 10 Gb 0.3 – 50 Gb/run

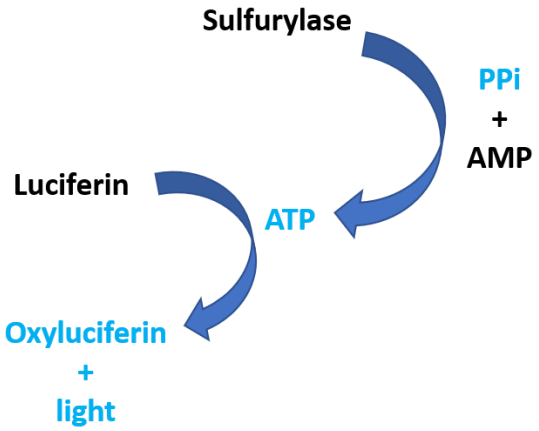
Pacific Biosciences (Third)	Sequencing-By-Synthesis	PacBio RS II Sequel	Up to 6 hours/cell Up to 20 hours/cell	Variable	0.5 – 1Gb 5 – 10 Gb
Oxford Nanopore Technologies (Third)	Sequencing-By-Translocation	SmidgION Flongle MinION GridION X5 PromethION	NA NA 2 days 2 days 2 days	Variable	Use with Smartphone Single use 10 – 20 Gb/run Up to 100 Gb/run Up to 12 Tb/run (at full capacity)

***** Data relevant to read length, run time and throughput specific to each platform were obtained from the corresponding manufacturer's website, product literature and allseq.com

Figure 1.2 454 technology. During sequencing the DNA strand, the addition of GTP results in the conversion of GMP and the inorganic pyrophosphate (PPi). The sulfurylase enzyme uses PPi to convert AMP to ATP which is then used by luciferase enzyme to produce oxyluciferin and light. This chain reaction resulting in generation of light signal is used to identify the corresponding bases added by the polymerase enzyme.

AGTACGTCTGAGACGTGA

GCACT



and G and this cycle is repeated until sequencing is complete. During sequencing, the incorporation of unlabeled nucleotides release pyrophosphate that is first converted to ATP and is then subsequently used by luciferase enzyme to generate oxyluciferin and photon as by-products (Figure 1.2). The photons are captured by the charged-coupled device at the bottom of the fiber-optic plate and the release of the photons following each nucleotide is converted to a corresponding base signal.

The initial release of the 454 GS 20 offered read lengths close to 100 bases but increases in number of wells per plate accompanied by subsequent improvements resulted in higher throughput. The GS FLX and Titanium series instruments produced longer read lengths close to 700 bases and a throughput of 0.7 Gbp per run and an accuracy of 99.9% and a faster turnaround time of less than a day [Rothberg & Leamon, 2008; Heather & Chain, 2016; Liu et al., 2012]. One of the main drawbacks of the 454 method is its challenge in sequencing homopolymer repeats. A linear increase in signal is observed for short repeats but as the length increases, broadening of signal occurs resulting in ambiguous basecalling [Quince et al., 2009; Margulies et al., 2005]. In addition, a small fraction of templates within each bead undergoes asynchronous sequencing that causes some of these templates to either lag-behind or go-forward due to insufficient, or the presence of, residual nucleotides, respectively [Margulies et al., 2005]. While the 454 technology offered longer read length and high quality, the throughput was lower than other NGS technologies and is more expensive on a cost per base basis. As a result, this technology was discontinued in mid-2016 [Liu et al., 2012; GenomeWeb, 2018].

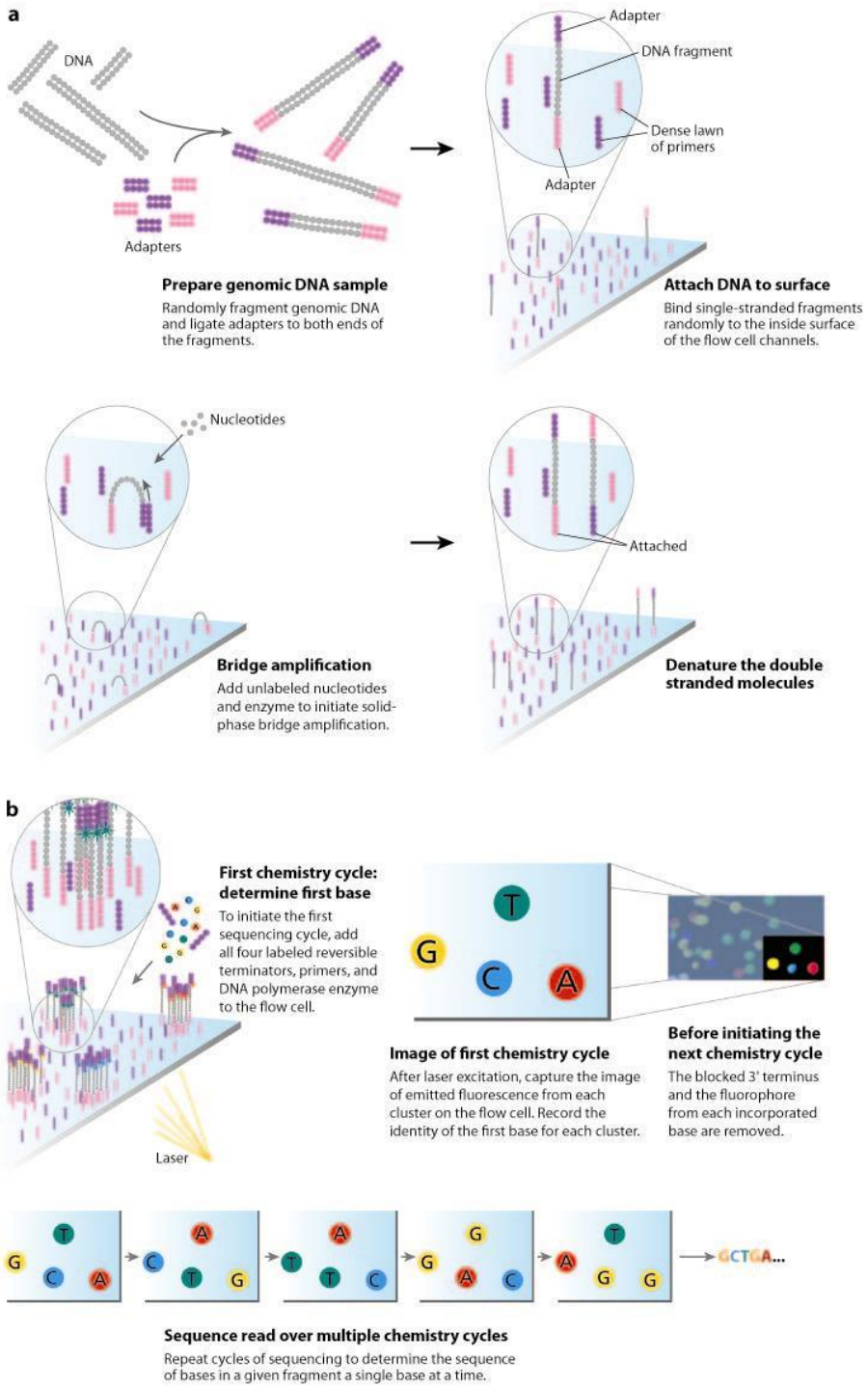
Illumina Platform

The short-read NGS technology offered by Illumina was first developed by Solexa in 2006 and later bought by Illumina [Liu et al., 2012]. The Illumina platform adopts a SBS approach and uses bridge amplification to generate clonal copies of target DNA molecules for sequencing. The library preparation in this method involves attaching platform specific adapters to each ends of

the target and is denatured and bound to flow cell containing complementary adapter sequences. The target DNA molecules are loaded on to flow cells containing complementary adapters and the templates are extended and copied by the DNA polymerase. The original library molecules are removed and clonal copies of each molecule are generated by a process called bridge amplification during which the single-stranded DNA binds the nearest adapter sequence that is complementary to its free 3'-end of the target by forming a bridge-like structure (Figure 1.3). DNA polymerase amplifies this template through the bridge and about 1000 clonal copies are generated in his process for each individual molecule in about 35 rounds of PCR amplification. The short size of the DNA template generates clusters in the near vicinity and about 800 to 1000 K clusters per mm^2 can be generated before sequencing [Buernas & den Dunnen, 2014]. In the next step, enzymes, primers and reversible terminator fluorescent nucleotides are supplied, and the presence of the terminator prevents the polymerase from further extending the chain thus allows images to be captured from all clusters in parallel. This cycle is repeated until sequencing is complete and the number of cycles used for sequencing determines the number of bases obtained as each cycle corresponds to image capturing associated with a base. Multiplexing allows the ability to sequence hundreds of samples together in a single run and involves attaching sample-specific barcodes in the amplification step. Once sequencing is complete, de-multiplexing allows the reads to be assigned to individual samples specified by the barcodes.

This Illumina technology is widely adopted by researchers worldwide and offers a vast array of instruments to customize sequencing experiments depending on the individual needs of research. With its initial release of the Illumina Genome Analyzer (GA), a throughput of approximately 1G per run and a read length of 36 bp was achieved. Further improvements in the GA increased the throughput to 50 Gbp per run and with the GA IIx instruments, even higher throughput of 85 G per run was achieved. With the development of paired-end chemistry, longer read length can be obtained as the template can be sequenced from both ends as opposed to

Figure 1.3 Illumina sequencing technology. A) The target DNA is sheared and ligated with illumine specific adapters. The adapter ligated DNA is bound to the flowcells through complementary adapter sequences and undergoes bridge amplification to generate clonal population of template DNA. B) During sequencing, the fluorescently labeled nucleotides are supplied one at a time. Upon laser excitation of flow cell following each addition, the color of emitted by the fluorescent group is used to identify the corresponding DNA base.



Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

only one end in the single end chemistry [Liu et al., 2012]. Currently, Illumina offers the HiSeq X, an ultra-high-throughput machine, that is capable of sequencing complete genomes to atleast 30X coverage and costs less than $1000 [Levy & Myers, 2016]. The flow cell lanes in the HiSeq X contains billions of nanowell like patterned structure that limits the clonal amplification of individual templates to each well and are spaced evenly.

Illumina offers consistent read accuracy of over 98% and a slight decrease in base quality is observed as read length increases [Liu et al., 2012; www.illumina.com]. The high-end instruments such as HiSeq 2500 and 2000 requires longer run time of between 5 and 10 days but the introduction of NovaSeq 6000 decreased the run time to less than 2 days [Table 1.1]. The longer times with HiSeq instruments are due to the high-quality imaging process following the addition of each base. To reduce the total sequencing time, Illumina developed other instruments such as MiSeq and NextSeq that require less time (less than a day) and are suitable for applications requiring less sequence coverage. The run times are shorter for the Miseq and NextSeq because the flow cells for each instrument contains one and two lanes respectively and the time associated with imaging is far less compared to HiSeq. In addition, the NextSeq uses a two-color imaging process that identifies four bases from only two images taken per cycle.

Ion Torrent Platform

The Ion Torrent platform was commercialized in 2011 and the chips used by Ion Torrent platform uses semiconductor technology to sequence DNA templates and is similar to the 454-technology [Rothberg et al., 2011]. This platform utilizes the SBS method and relies on measuring the change in pH released during the incorporation of nucleotides. The DNA templates are ligated with adapters on both ends and captured on beads. Using emulsion PCR, a clonal pool of amplicon is generated for each individual molecule bound on bead and loaded on Ion Torrent chips containing micro-fabricated wells use CMOS semiconductor technology to sense change in release of protons. Each enzymatic incorporation of nucleotides results in the release of proton

and to distinguish bases on template at each position, the nucleotides are added to the chip in a sequential manner. Since there is no fluorescence-based imaging involved, the run times are much shorter, in the order of hours, compared to other imaging-based methods [Buermans & den Dunnen, 2014]

The initial release of Ion-314 chip contained only about 1.2 M wells and produced 10 Mb output. By increasing the surface area and the number of sensors, over 650 M wells could be fabricated on Proton-II chips representing over a 500-fold increase [Rothberg et al., 2011; Buermans & den dunnen, 2014; Merriman et al., 2012]. With the introduction of 500 series Ion chips, read lengths between 200 and 600 bases can be obtained but the quality score drops as read length increases [Rothberg et al., 2011; Liu et al., 2012; Thermofisher]. Since the Ion Torrent uses non-optical method and unmodified nucleotides, all four nucleotides need to be supplied in a sequential order at each cycle and washed away. Similar to the 454 platform, the presence of excess or insufficient nucleotides can result in loss of synchronicity and on average, 1% of the molecule undergo asynchronous synthesis [Buermans & den Dunnen, 2014]. Presence of homopolymer runs and AT rich regions pose challenge in the Ion Torrent platform but no bias was observed in GC rich region. The smaller repeats usually produce an intense signal but as repeat length increases to 5 or more bases, the error rate increases to about 3.5% due to lack of linear increase in signal corresponding to the repeat length [Merriman et al., 2012; Quail et al., 2012].

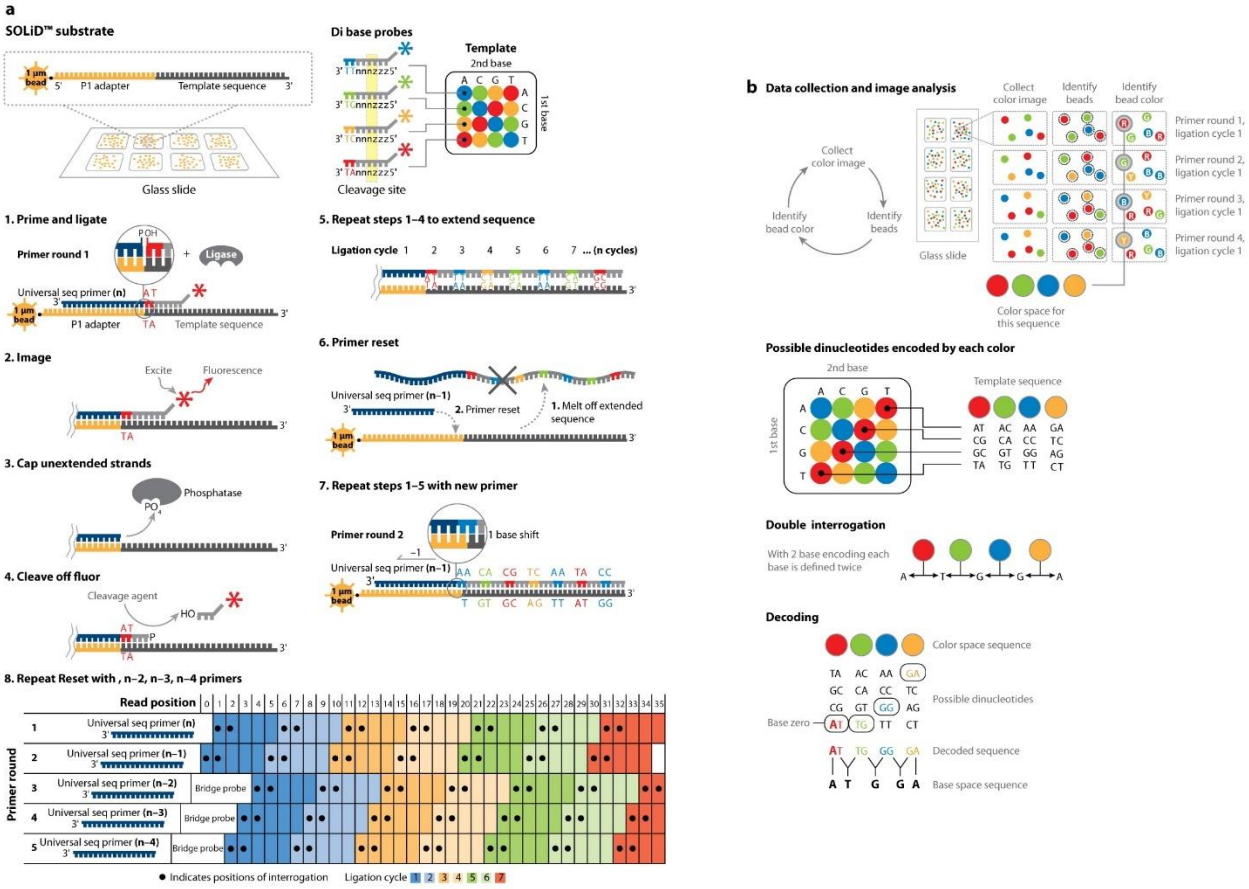
SOLiD Platform

The Sequencing by Oligonucleotide Ligation Detection (SOLiD) platform offers the ability to sequence DNA molecules by a different approach called sequencing-by-ligation (SBL). This technology was first commercialized in 2006 and uses DNA ligase to sequence targets that is different from other NGS platforms that uses DNA polymerase for sequencing [Liu et al., 2012]. In the SOLiD approach, the target DNA is ligated to adapters and are then bound to beads

containing complementary adapter sequence. After generating clonally amplified products using emulsion PCR, the beads are deposited on to a glass slide and supplied with a universal primer, DNA ligase, and four fluorescently labeled probes for ligation. The fluorescent label is covalently attached to the last nucleotide of the 8-mer probe and each fluorescent probe identifies four of the 16 possible di-nucleotide sequences. A color-space corresponding to the first two positions of the probe is imaged during each ligation step and two di-base colors are used to identify a single base position. For example, the base A in the following sequence GTGATGC will be identified by both GA and AT probes. Following image capture, the probe is cleaved between positions 5 and 6 to remove the fluorescent label and this cycle of ligation, detection and cleavage is repeated for a total of five rounds with each round starting with a universal primer that is offset by one nucleotide [Mardis 2008; McKernan et al. 2006].

The SOLiD method produces yields of over 2 billion reads per run with read accuracies up to 99.94% as each base position in the template is sequenced twice. This platform offers both single and paired-end sequencing can read upto 120 bases with each mate pair reading upto 60 bases [Allseq.com]. The SOLiD platform provides relatively inexpensive sequencing with the cost per base at approximately one-tenth of other conventional methods [Liu et al., 2012; Shendure et al., 2005]. Since the SOLiD technology relies on DNA ligase to perform sequencing, templates containing palindromic sequences are not suitable because the formation of hairpin structure can impede ligation of probes and thus sequence identification [Huang et al., 2012]. The run times are longer for the SOLiD platform compared to other platforms and require one and two weeks for single and paired-end sequencing [Liu et al., 2012]. The sequencing run generates about 4 Tb of raw data and requires more computational resources to perform analysis.

Figure 1.4. SOLiD sequencing. Panel A outlines the sequencing steps used in SOLiD sequencing. The DNA library containing SOLiD specific adapter is bound to the beads and deposited onto the glass slide. The sequencing starts with the binding of universal primer to the template followed by ligation of fluorescent di-base probes. Laser excitation allows the identification of ligated di-base probe and the fluorescent dyes are cleaved off before ligating next probe. After completing one full-round of sequencing, the extended strand is removed and the new round of sequencing starts with another universal primer that is reset by one base and the primers are rese for five rounds. Panel B shows the data analysis following imaging in each round of sequencing. The fluorescent colors for each probe represents a di-base and by interrogating the color space for each round of ligation, the DNA sequences can be identified.



Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

DNA Nanoball Technology

The DNA nanoball (DNB) technology employs a ligation-based approach that uses combinatorial probe-anchors for sequencing, similar to the SOLiD platform. This technology generates a circular DNA nanoball that contains the target DNA of interest separated by adapter sequences [Drmanac et al., 2010]. The first step in the library preparation involves adaptor ligation to each ends of the target followed by circularization to close the ends. This cycle of ligation and circularization is continued for three additional cycles such that adaptor sequences intersperse template DNA by about 70 bases that can be sequenced. The Phi29 polymerase is used to replicate the DNB mediated by rolling circle amplification and the palindromic regions present in adapter sequences promote the assembly of linear sequences in nanoball-like structures. The DNBs are adsorbed electrostatically onto a silica plate, similar to a flow cell, to form a patterned nanoarray structures that contains ~350 million spots. Using the combinatorial probe-anchor ligation (cPAL) chemistry, fluorescently labeled probes are imaged after each round of ligation and the bases are identified. This method allows both single and paired-end sequencing from the four adapter regions and up to 70 bases can be sequenced [Porreca, 2010]. The average sequencing cost per genome using this method was ~$1,400 and the raw sequencing accuracy compares to other methods. This method is easily scalable to large sample volume and by inserting additional adapters, up to 120 bases can be sequenced [Drmanac et al., 2010].

Third Generation Sequencing

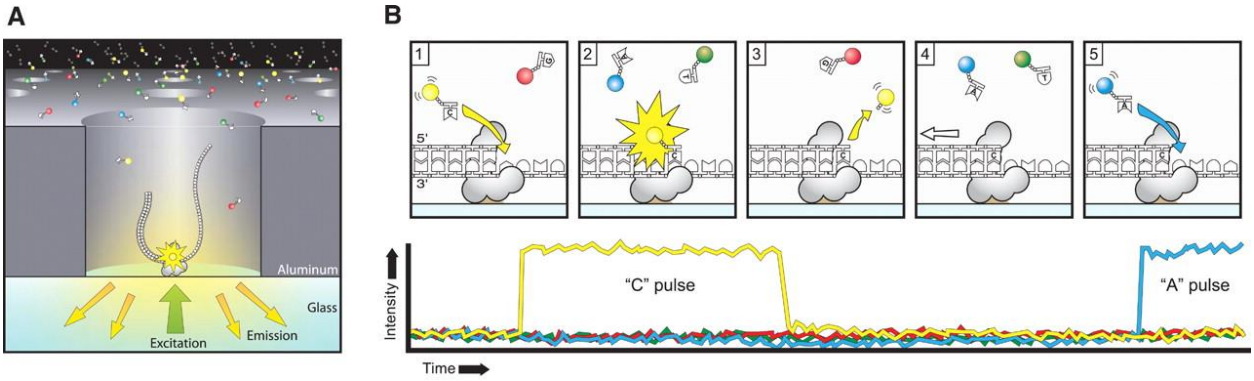
The third-generation technologies for sequencing DNA templates are clearly distinguished from their second-generation counterparts by their ability to produce relatively longer reads. The template molecules can be sequenced in real-time without the need for clonal amplification. The real-time sequencing utilizes the processive nature of enzymes thus eliminating the amount of time associated with cycling between incorporation of nucleotides which is usually 1 hour for the HiSeq2000 and 5 mins for the MiSeq [Buernmans & den Dunnen 2014]. Although TGS methods

are commercially available only recently, these concepts have been described earlier and were under active development [Eid et al., 2009, Kasianowicz et al., 1996]. Several novel third generation sequencing methods have been proposed that uses electron microscopy, nanopore, fluorescence resonance energy transfer and transistor-based approaches, PacBio and Oxford Nanopore Technologies (ONT) are currently commercially available for sequencing [Schadt et al., 2010].

PacBio Sequencing

The single molecule real-time (SMRT) technology developed by Pacific Biosciences sequences template molecules by measuring the incorporation of nucleotides as the DNA polymerase synthesizes its growing strand. The SMRT method uses zero mode waveguides (ZMW) that allows the imaging to be confined to a smaller volume of 10^{-21} liter where the enzyme is immobilized enabling video capture of individual nucleotide synthesis [Eid et al., 2009]. The SMRT cells use a mutant Phi29 polymerase that retains the properties of high processivity, strand displacement, no GC bias and low error rates making it an attractive choice for sequencing [Buermans & den Dunnen, 2014]. While the SMRT sequencing is conceptually similar to the reversible terminator method, it uses terminally linked phosphonucleotides that allows real-time continuous monitoring of nucleotide incorporation without the need for terminator cleavage. The library preparation involves ligation of SMRT bell adapters to each end of the double stranded templates resulting in the generation of single stranded circular molecules that is loaded onto ZMWs for sequencing. Because the template is circular in nature, the enzyme can sequence the template more than once resulting in continuous long reads (CLR). The SMRT bell adapters from CLR can be trimmed to generate subreads and these subreads can be used to generate a circular consensus reads (CCS) [Rhoads & Au, 2015]. The average read length obtained from this platform increased with improvements and it currently gives an average of over 15 kb with fewer reads spanning over 100,000 kb in length [www.pacb.com].

Figure 1.5. Single Molecule Real Time (SMRT) sequencing technology. A) The ZMW containing a bound DNA polymerase in an individual well is shown. The dNTPs are shown in different colors that are in constantly flowing in and out of each well. As DNA polymerase synthesizes DNA, florescence corresponding to each nucleotide is captured as a light pulse in the movie of a movie. B) The fluorescence signal corresponding to the addition of Cytosine followed by Adenosine is shown. As each nucleotide is incorporated during sequencing, fluorescent signals corresponding to other nucleotides are only low.


A
Aluminum
Glass
Excitation
Emission
B
1
2
3
4
5
5'
3'
Intensity
"C" pulse
"A" pulse
Time

Although third-generation technologies can generate longer reads, they suffer from higher error rate and lower throughput compared to other NGS methods. The error rate for PacBio ranges between 11% and 15% but as errors are randomly distributed throughout the read, accuracy can be improved by generating consensus read provided multiple subreads are available for each individual template [Rhoads & Au, 2015]. The PacBio shows relatively small bias towards GC rich regions but it can sequence extremely AT rich regions [Quail et al., 2012]. The throughput is affected by the life of polymerase and the runtime for each ZMW cell [Buermans & den Dunnen, 2014]. The RS II SMRT cells contain approximately 150,000 ZMWs, one third of which produce reads resulting in a throughput of up to 1 Gb data during a 4-hour run [Eid et al., 2009; Rhoads & Au, 2015]. The recently released Sequel system contains 1 million ZMWs that has the potential to deliver up to 20 Gb per run. While the Sequel system offers a higher throughput compared to RS II, it is still inferior to other NGS technologies. Both the RS II and Sequel offer flexibility in the runtime needed per SMRT cell and the number of SMRT cells is modular and can be increased as needed for the experiments [www.pacb.com].

Oxford Nanopore Technology

The ONT utilizes a sequencing-by-translocation (SBT) approach whereby sequencing is performed as the template strand passes through a nanopore embedded into a membrane array. An ionic gradient established between a synthetic membrane and the application of voltage allows the flow of ions through the pores generating a small electric current and as DNA passes through the pore, a characteristic change in the flow of electric current corresponding to each base is observed that is the basis of nanopore sequencing [Ip et al.,2015]. A mutant form of *Escherichia coli* CsgG protein is used as pores in the latest version of flow cells which has improved accuracy, higher throughput, higher pore stability and increased translocation speed of 250 b/s compared to the previous versions [Carter & Hussain, 2017]. In addition to the membrane pores, a tether to

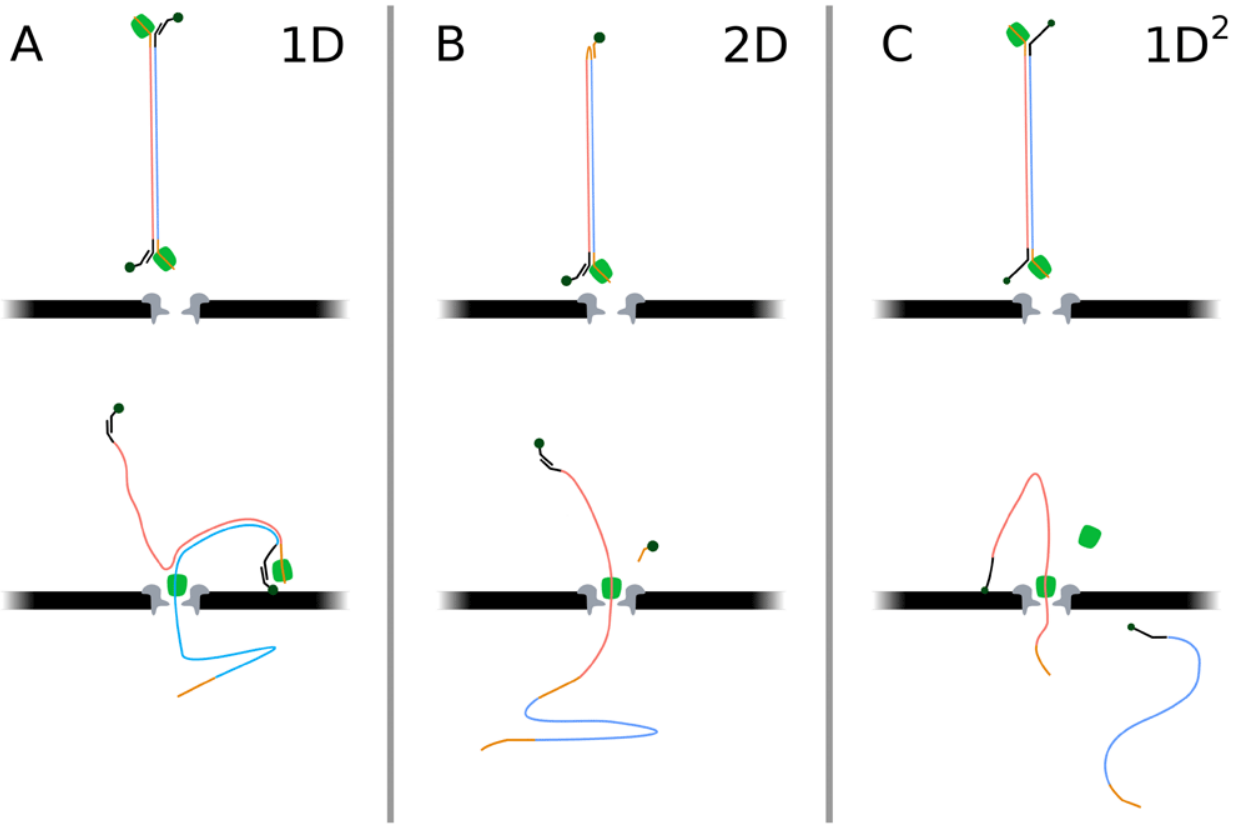
bring the target DNA in close proximity to the pore and motor proteins to unwind the target are necessary for sequencing [Ip et al., 2015].

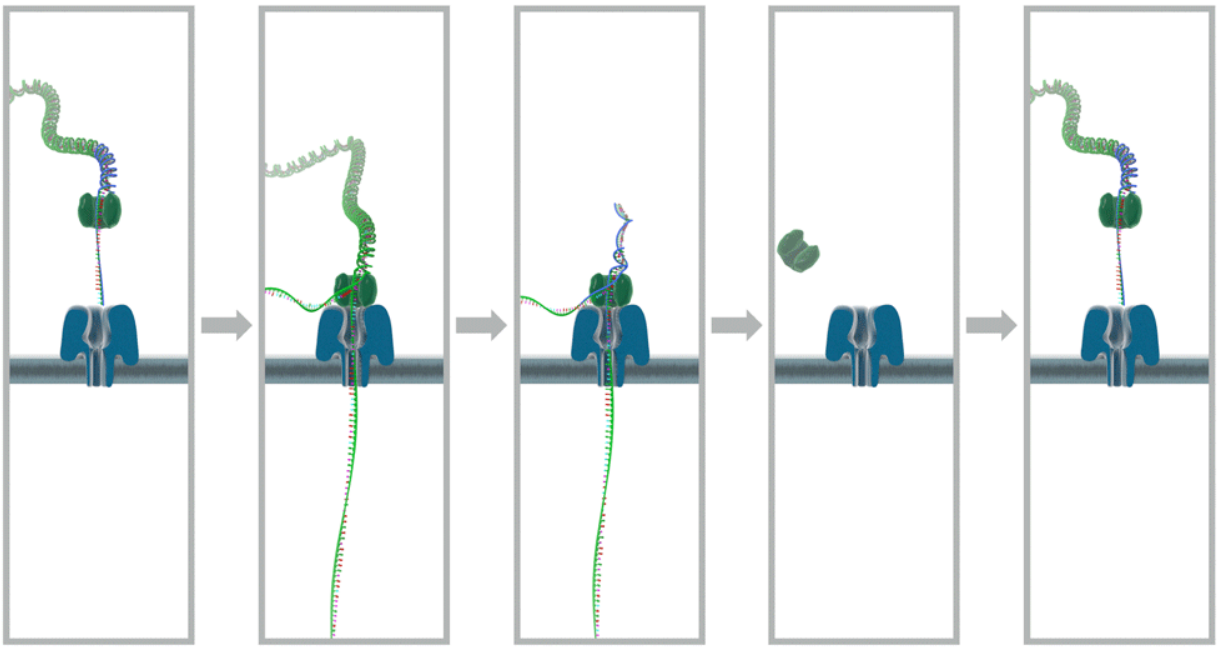
The 2D library preparation involves taking the double stranded DNA through end repair followed by A tailing, ligation of leader and hairpin adapters and tethering the library for sequencing. This earlier protocol version provided the ability to covalently link the two strands of the target with hairpin adapters that when sequenced can result in template (sequenced first) and complement (sequenced second) 1D reads from each strand. The sequence information separating the template and complement reads are identified by the basecaller when a characteristic current signal corresponding to an abasic site present in the hairpin is recognized [Quick et al., 2014]. A high quality 2D read can be generated for each target sequence provided each of the 1D reads are present. The latest ONT protocols with R9 chemistry eliminates the need to covalently link two strands significantly thus decreasing the sample preparation time to only 10 minutes. Despite the strands not being linked, each strand can be sequenced separately resulting in reads with higher accuracy [Lannoy et al., 2017].

A key advantage of MinION is that its smaller size and weight of only about 100 g makes the sequencer portable. The ONT platform is different from PacBio in its ability to generate ultra-long reads of over 0.8 Mb [Jain et al., 2018]. In addition to MinION, Oxford Nanopore offers GridION X5 and PromethION that provides modular control over the number of flowcells and higher throughput. Both MinION and GridION contains 512 pores per flow cell but PromethION contains up to 3000 pores per cell with the ability to sequence 48 flow cells at a time [nanoporetech.com]. The throughput is dependent on the number of active pores available at any time and the number of pores per flow cell is increasing with new improvements in technology.

Since its first access to the research community in 2014, improvements in sequencing chemistry and basecalling have contributed to an overall increase in accuracy, read length and data throughput but there is a considerable variability between experiments that needs to be optimized [Ip et al., 2015; Jain et al., 2018]. With R9.0 sequencing chemistry, the median accuracy

Figure 1.6. Oxford Nanopore Sequencing Technology. Panels A, B and C show the different sequencing chemistries used in ONT technology. Both 1D (panel A) and $1D^2$ (panel C) chemistries sequence only one strand of the double stranded (ds) DNA but the 2D chemistry (panel B) covalently links the two strands of dsDNA and sequences one linear molecule. The motor protein is shown in green color. Panel D shows the sequencing of DNA strands using nanopores. The motor protein first unwinds the DNA strand and a single strand is first passed through the nanopore. Once the DNA is completely passed, the motor protein dislodges from the nanopore and the nanopore is ready to sequence another DNA template bound by motor protein (Figure obtained from de Lannoy et al., 2017 under open access from Creative Commons Attribution License).


A
1D
B
2D
C
1D2



is between 89% and 94% but with R9.4 chemistry combined with improved basecalling and faster sequencing (450 b/s), the consensus read accuracy reached 99.75% [Wick, et al., 2018]. Sequencing experiments with R7 and R9.0 chemistries showed bias against GC rich region [Jain et al., 2018; Laver et al., 2015] but the R9.4 chemistry did not show this bias [Carter & Hussain, 2017]. The ONT's Metrichor basecaller posed challenges to call homopolymer repeat sequences as it under and over represented AT- and GC-rich regions respectively. The development of Scrappie addressed this bias and showed better performance when compared to Metrichor and comparable to Illumina sequencing [Jain et al., 2018]. Scrappie is currently under active development and with the ONT technology still in its infancy, rapid improvements are expected to occur that can address some of these bottlenecks [Brown & Clarke, 2016; Wick et al., 2018].

Other Long-read Technologies

Illumina's Moleculo technology was based on the LR-seq method that utilizes short-read sequencing data to reconstruct synthetic long-reads (SLR) [McCoy et al., 2014]. In this method, long single stranded cDNAs are synthesized and using adapter sequences present in both ends, double stranded molecules are generated. In a 384-well plate, between 10^3 and 10^4 molecules are loaded onto each well and the long DNA present in each well is further fragmented into short reads that can be sequenced on the Illumina platform. Barcodes specific to each well are attached to the short DNA fragments and the samples are pooled together for sequencing at high depth. The reads are demultiplexed and assigned to individual wells using the barcodes and assembled to SLRs in a well-specific manner. This method relies on the assumption that the probability of two isoforms originating from same gene ending up in the same well is very low. The read accuracy is characteristic of Illumina platform because the long reads are assembled from short reads. The SLRs generated by this approach have been used to haplotype genomes that resolves upto 99% of SNVs [Kuleshov et al., 2014]. By using as few as 1000 cDNA molecules per well in

a 384-well format, molecular co-association between distant exons has been studied in the human genome [Tilgner et al., 2015].

Base4 is a UK-based startup that is currently developing a sequencing technology based on the pyrophosphorolysis method [www.base4.co.uk]. The double stranded template DNA molecules are passed through a tube-like structure where it undergoes pyrophosphorolysis, a molecular process that catalyzes the release of terminal nucleotides from the DNA chain in the presence of pyrophosphate. The released nucleotides are captured into micro-sized oil droplets and a proprietary Cascade Reaction occurs within each droplet that produces bright fluorescent signal that are then converted to individual bases. The order by which the droplets are basecalled generates the complete sequence of template DNA. This technology claims it is capable of sequencing upto 1Mb per second and low systematic error [Eisenstein, 2015; www.base4.co.uk].

Direct RNA sequencing

In addition to sequencing DNA molecules in a massively parallel manner, two additional methods to directly sequence RNA molecules were developed. The first method uses an SBS approach and in this method, the RNA molecules containing poly(A) tail is bound to poly(T) sequences covalently attached to the surface of beads [Ozsolak et al., 2009]. For RNA molecules that are non-poly adenylated, the A-tails can be added with E.coli poly(A) polymerase and by adding dideoxy nucleotide in the reaction mixture, the extension of A-tail can be stopped and the tails can be grow to desired length. In the next step, sequencing is performed with DNA polymerase and at each step of nucleotide addition, the fluorescence is measured that corresponds to the one of the four bases. While this method is promising and can be used to directly sequence RNA molecules, the mean read length obtained from the prototype sequencer is only 28 bases long and maximum read length of up to 60 bases were obtained. The error rate is also high around 4% and it takes three days to complete 120 cycles of sequencing [Ozsolak & Milos, 2011].

The second method for direct RNA sequencing is developed by ONT that uses the sequencing-by-translocation method with the help of an engineered nanopore protein. This method is principally similar to the DNA sequencing offered by ONT platform and the library preparation step involves the addition of an adapter sequence to the 3'-end of the RNA [Garalde et al., 2018]. A motor protein that is bound to the adapter unwinds the RNA through the nanopore in the membrane and the change in electric current due to RNA translocation is converted to corresponding bases and thus the RNA sequence. Currently this method can be used to sequence RNA molecules that contain poly(A) tails but custom adapter sequences complementary to the 3'-end of RNA can be designed to sequence RNA that lacks poly(A)-tails. By using a modified synthetic RNA containing N6-methyladenosine and 5-methylcytosine, the characteristic change in current profile observed by the modified bases compared to unmodified bases facilitates detection of modified RNA bases and this method is promising to detect other RNA modifications as well [Garalde et al., 2018].

Bioinformatic tools to analyze high-throughput sequencing data

As high-throughput sequencing (HTS) technologies emerged to deliver large volume of sequence data in a single run, the need for memory efficient alignment tools increased because the tools available at that time were not efficient at aligning short reads and required significant computational resources [Li et al., 2010]. This gave rise to the development of number of different tools that work at each and every stage of the alignment process starting from the initial quality control to the final steps of data visualization [Pabinger et al., 2012]. The first step in the NGS pipeline involves data collection followed by quality control that analyzes the read quality and trims low quality regions. The pre-processed reads are then aligned to the reference genome, transcripts are assembled and the expression of each genomic region is measured in the form of normalized read counts that can be used for differential expression analysis for genes and transcripts among different treatment conditions.

Due to the availability of various bioinformatic tools, many groups have shown interest in evaluating the performance of these tools and benchmarked them on the basis of their ability to align large numbers of reads, the computational time required, and trade-offs between read error and alignment accuracy [Li & Homer, 2010; Hatem et al., 2013; Ruffalo et al., 2011; Yu et al., 2012]. Bowtie, one of the most commonly used aligner in the NGS field was built on Burrows-Wheeler Transformation algorithm. Bowtie uses a novel indexing strategy to reduce the computational time required and achieves this ultra-fast performance at the expense of missing some read alignments when the query contains multiple mismatches [Langmead et al., 2009]. The initial release of Bowtie did not allow gapped alignments and since the RNA-seq reads usually came from spliced RNA molecules, aligning these reads to genome would result in gaps between exons and Bowtie does not align these reads to these regions resulting in unmapped reads. To address this issue and to increase the fraction of aligned reads, TopHat was developed [Trapnell, et al., 2009]. TopHat is a splice-aware program that aligns reads in two steps. The first step uses Bowtie to generate alignments without allowing any gaps and the exon junctions inferred from this step are used in the second step to align previously unmapped reads. The Spliced Transcripts Alignment to Reference (STAR) tool also supports longer reads and gapped alignments that span exon junctions [Dobin et al., 2013]. For transcriptome profiling experiments that involve performing differential expression analysis, the raw alignment data must be converted to a useful measure. Cufflinks [Trapnell et al., 2010] tool assembles transcript from the alignment data and estimates the abundances of genes and transcripts and reports them in units of Reads per Kilobase per Million mapped reads (RPKM), an expression metric that is normalized to gene length and sequencing depth, can be used to compare different samples. There are several other tools that can be used to perform differential expression analysis but the type of normalization used can affect the number of differentially expressed genes [Dillies et al., 2012] and differences in performance have been observed between various differential expression algorithms [Teng et al., 2016].

The alignment tools described above all requires a reference genome upon which the reads can be aligned but in the absence of such reference, *de novo* sequencing assembly can be performed [Trapnell et al., 2012]. Since the RNA-seq data provides expression information about individual transcripts, the coverage is not uniform across all transcripts in a gene and the *de novo* transcriptome assembly can be a challenging task. Trinity, a modular tool, assembles *de novo* transcripts by building contigs and de Bruijn transcript graphs and performs well in terms of reconstructing full-length transcripts compared to other assemblers [Grabher et al., 2011]. Bridger is another *de novo* assembler that uses concepts from both Cufflinks and Trinity to reduce the false positive rate compared to Trinity [Chang et al., 2015].

The long-read technologies offered by PacBio and ONT have relatively high error rates compared to short read technologies but they are capable of providing long, contiguous reads. The higher error rate inherent to these technologies can result in a large fraction of reads that are unmapped when using these short-read aligners because they are built to handle short reads of very high quality. LAST and Graphmap are two tools that perform well in aligning long reads compared to other available tools [Sovic et al., 2016]. Both of these tools use a seed-and-extend approach and Graphmap allows gaps in seed regions that can be extended.

Applications of Sequencing Technologies

The sequencing technology has broad applications in many research and clinical setting and with an arsenal of different sequencing platforms available to generate both long and short reads, a vast array of techniques have been developed [Illumina handbook, 2018]. The information coded in the genomic DNA is a treasure trove containing a vast amount of information and below, I will discuss the five broad sequencing applications in detail.

Genome Sequencing

One of the most widely used applications of NGS technology is the genomic sequencing to build genome assemblies and resequencing to identify minor structural variations such as SNPSs and indels. Since the completion of first human genome sequencing project, the number of DNA bases in the GenBank repository steadily increased reaching over 2 trillion bases [Shendure et al., 2017] and the understanding and functioning of our genome has vastly improved. The short-reads obtained from NGS platforms can be used to assemble genomes either *de novo* or to a reference sequence. Earlier approaches to sequence the human genome were based on a shotgun approach where the overlapping reads are used to build contiguous sequences (contigs) which are then assembled into genome scaffolds. The scaffolds contain gaps between contigs that mainly arise due to low coverage sequence information due to repetitive regions and these regions can extend on the order of megabases [Eichler et al., 2004]. The first human assembly contained 321 interstitial gaps in euchromatic regions but it has been reduced to only 164 gaps in GRCh37 build due to a combination of improvements in sequencing technology, library preparation methods and the availability of efficient bioinformatic tools providing us with a near complete genome [Eichler et al., 2004; Chaisson et al., 2015]. In the past decade, approximately 13,000 prokaryotic, eukaryotic and viral genomes have been sequenced with over 9,000 complete assemblies are available [Kitts et al., 2015].

Genome resequencing experiments are used to identify minor SVs associated with individual genetic variation. The high-quality and high-throughput features of NGS technologies makes it aptly suited to resequence individual protein-coding genes or whole exomes. A population level resequencing targeted towards *ANGPTL4,* a adipokine gene involved in lipid metabolism identified a rare variant in European Americans significantly associated with high levels of HDL [Romeo et al., 2007; Topol & Frazer, 2007]. The 1000 Genomes project has catalogued extensive genetic variation between individuals from various populations [The 1000G consortium, 2012]. Similarly, large-scale genome sequencing efforts are currently underway in

many other parts of the world towards cataloging genetic variations among populations [An, 2017].

Transcriptome Analysis

The transcriptome contains a diverse array of RNA molecules and refers to the set of all transcripts, both coding and non-coding, present in a given cell or tissue at any point in time. Although the human genome contains 3 billion base pairs, it is limited in its size but the coding potential of the genome measured by the transcriptional output is astonishingly complex [Pan et al., 2008; Wang et al., 2008; Djebali et al., 2012]. Regulation occurs at multiple levels, both pre- and post-transcription and the mechanism of alternative splicing (AS) controls how transcripts are spliced together to generate multiple isoforms [McManus & Graveley, 2011]. Our understanding of the transcriptome complexity has vastly improved over the years and NGS methods have determined that over 95% of genes with multiple exons undergo alternative splicing, tremendously increasing the coding potential of the genome [Nilsen & Graveley, 2010]. In addition to the complexity exhibited by global transcriptome, fascinating examples of complexity present at individual gene levels are also present that can potentially generate hundreds of thousands of different isoforms [Graveley, 2001, Park & Graveley, 2007]. The *Down syndrome cell adhesion molecule 1* (*Dscam1*) gene in *Drosophila melanogaster* is known to undergo the most extensive alternative splicing of any gene and can generate 38,016 isoforms from 95 cassette exons which far exceeds the total number of number of genes present in the entire organism [Graveley, 2005].

Expressed sequence tag (ESTs) [Adams et al., 1995] containing partial complementary DNA from messenger RNA, is a digital read out of expressed genes were used initially to identify the transcribed regions. But the lowly expressed genes are difficult to detect with this method and the high cost of capillary sequencing limited its ability to catalogue the transcriptome in a comprehensive manner. Other tag-based methods such as serial analysis of gene expression (SAGE) and cap analysis of gene expression (CAGE) have addressed these drawbacks

associated with ESTs and enabled studying the expression of low abundance genes, and profiling the transcription start sites and alternative promoter usage. Microarrays that utilize probes can detect changes in expression levels in individual genes but the low signal to noise ratio and the tedious design of probes from the 3' end of genes limit its utility to study genes with large numbers of isoforms. The splicing-sensitive microarrays have been used to study isoform specific expression changes but the design of exon junction probes to identify all possible isoforms in the global transcriptome can easily turn to become a complex task [Lee & Roy, 2004]. For example, to study *Dscam1* isoforms using microarray, custom probes need to be designed for the 93 alternative exons to detect expression changes from different tissues [Neves et al., 2004]. The advent of NGS technologies have addressed these drawbacks and the number of reads that map to a given gene becomes the digital readout of expression. These methods have been utilized to study the dynamics of the transcriptome and catalogue the extensive variation in alternative splicing between different tissues and developmental stages in greater detail. The ENCODE and modENCODE projects have profiled the transcriptome of different cells, tissues, developmental stages and environmental perturbations in worms [Gerstein et al., 2010], flies [Cherbas et al., 2011; Brown et al., 2014; Graveley et al., 2011] and humans [Djebali et al., 2012]. The RNA-seq data generated from these studies have identified many novel transcripts, both coding and non-coding, that were previously unannotated, discovered the expression of new alternative exons and novel RNA editing sites.

The application of NGS methods are not limited to characterizing the changes in the global transcriptome at the population or tissue levels but individual cells can also be studied. Single cell transcriptomics relies on efficient capture of individual cells but following cell capture, the NGS platform specific library preparation follows that first synthesizes cDNA from single cells. Combining single-cell methods with NGS approaches offers a distinctive advantage to profile individual cells and can provide a snapshot of cellular heterogeneity within a homogenous cell population [Kolodziejczyk et al., 2015]. Our preliminary experiments to characterize 96 S2 cells

from *Drosophila* using the Fluidigm C1 chip identified the expression of different number of genes between cells and quite interestingly, the expression of *Dscam1* isoforms was limited to atmost one isoform per cell for a large fraction of the cells but none expressed more than one isoform (unpublished data). Many biologically important and interesting questions can be addressed using high throughput sequencing methods. The RNA-seq methods have provided the ability to map interactions between target RNA and its *trans*-acting RNA binding proteins (RNABPs) at a single nucleotide resolution [Van Nostrand et al., 2016] and to measure the affinity of RNA for different RNABPs [Lambert et al., 2014]. With the continuous development and adaptation of RNA-seq methods, the application of NGS in transcriptomics is far-reaching.

Epigenomics

Similar to RNA-seq, the DNA-seq experiments have enabled comprehensive characterization and identification of distinct molecular signatures that mark different genomic regions. The histone proteins that act as scaffolds to wrap genome into nucleosomes can undergo many different post-translational modifications [Bannister & Kouzarides, 2011]. The chromatin immunoprecipitation experiments enrich for DNA sequences bound by certain proteins or histones and before the advent of NGS, both quantitative PCR and southern blotting techniques were used to identify whether the captured fragments contain the regions of interest [Mardis et al., 2008]. The introduction of microarrays offered the ability to study genome-wide DNA-protein interactions using ChIP-chip approaches where the “chip” contains probes to simultaneously detect multiple regions captured from “ChIP” experiments (Chromatin Immuno Precipitation). By combining ChIP with NGS (-seq) methods, the drawbacks of the other approaches such as low throughput and low signal to noise ratio were addressed. The chromatin signatures from nine different human cell types was profiled using ChIP-seq experiments in the ENCODE project [Ernst et al., 2011]. This study revealed the identification of 15 different chromatin states from nine chromatin marks that corresponds to various genomic features such as enhancers and promoters using machine

learning approach. To detect methylated regions in CpG regions in the genome, the DNA can be treated with bisulfite that selectively converts unmethylated cytosines to uracil whereas the methylated cytosines are unaffected and are sequenced as C's.

Metagenomics

The field of metagenomics studies the genomic diversity of microbes in environmental samples and provides information about the composition of microbial community. Before high throughput sequencing methods were available, the microbes were studied primarily through culture methods and is often limited in its ability to study microbes that can be cultured and there are many bacteria that cannot yield to researchers attempt to culture [Amann et al., 1995]. The shotgun genome sequencing approaches and NGS methods allow the rapid identification of microbes and the ability to conduct comprehensive survey of microbial communities in a culture-independent manner [Tyson et al., 2004]. Metagenomics has been applied to study bacterial communities present in different ecosystems ranging from salt water to extreme conditions such as volcanoes [Oulas et al., 2015]. The coverage depth in sequencing-based methods can provide information about the degree of variability of individual organisms and the relative abundance of each species within a population. The study of the human microbiome is an important application of NGS methods in metagenomics that sheds light on the interplay and dynamics between microbial diversity and human health. The microbiome present in different regions of human body including distal guts [Gill et al., 2006], intestine [Eckburg et al., 2005] and other regions [Cho & Blasér, 2012] have been studied and led to the identification of microbes that novel and have not been cultivated before [Eckburg et al., 2005]. The long-read technologies have been applied in metagenomic studies to complement the shortcomings of short read technologies to assemble contigs into larger scaffolds and to build complete genomes and currently, long-read technologies are used to decipher the metagenomes in less diverse communities or dominated by a few microbes [Driscoll et al., 2017; Brown et al., 2017]. With the introduction of PacBio's Sequel and

ONT's PromethION, high throughput long read sequencing can potentially be applied to study even rare species in the metagenomes.

Clinical Applications

The development and availability of high-throughput sequencing technologies have tremendously changed the way how tot identify and diagnose deadly disease outbreaks. In the early nineteenth century, the H1N1 influenza outbreak wiped out almost 3% of the world's population with the lack of an understanding of how the disease spreads. But a similar pandemic that occurred in early twentieth century in 2003 due to Severe Acute Respiratory Syndrome (SARS) virus with a similar transmissibility rate to that of H1N1 resulted in far fewer fatalities [Lipkin, 2013]. The sequencing of the 30 kb RNA SARS genome occurred within a few weeks and the phylogenetic analysis of sequences revealed that this Coronavirus was only moderately related to other viruses in the same family and enabled the development of PCR-based assays for early detection [Marra et al., 2003; Rota et al., 2003]. With the advent of NGS technologies, sequences of pathogenic viral genomes have been obtained at an accelerated pace reducing the time from weeks to days [Briese et al., 2009]. The availability of portable sequencers even made the outbreak surveillance easier and much faster, within hours, by providing the ability to sequence in real-time directly in the field have been demonstrated recently with Salmonella [Quick et al., 2015], Ebola [Quick et al., 2016] and Zika [Faria et al., 2016] outbreaks.

In addition to surveilling outbreaks, NGS technologies have many implications in other clinical domains including genetic screening for diagnostic and treatment purposes. Genetic screening is an evaluation method to detect risks associated with certain disease phenotypes in asymptomatic individuals and such screening methods can have profound implications for the health of patients. Prenatal screening tests are used to detect fetal aneuploidies that uses a combination of test but the positive screening requires confirmation with invasive diagnostic procedures such as amniocentesis bears the risk of miscarriage [Pitukkijronnakorn et al., 2011].

The current non-invasive procedures of sequencing cell-free fetal DNA present in the maternal plasma and the NIFTY method has been demonstrated to identify trisomy in chromosomes 21, 18 and 13 [Jiang et al., 2012]. Gene panels target a small subset of genes to identify any changes with small indels and SNVs and have been widely used. In cancer diagnostics, the rapid decline in the cost of NGS technology makes genetic screening more readily available to clinicians who can offer to individuals at the risk of developing *de novo* mutations. Targeted sequencing of TP53 and BRCA1/2 from patient samples identified many novel point mutations and indels upto 16 nucleotides [Morgan et al., 2010]. A multi-gene panel study assaying 40 genes in breast cancer patients who tested negative for BRCA1/2 mutations identified 16 pathogenic variants in 9 different genes underscoring the importance of sequencing multiple genes as opposed to single genes [Kurian et al., 2014]. The application of whole exome sequencing (WES) and WGS approaches as a clinical diagnostic utility is emerging and is better suited to study complex disease conditions such as neurodevelopmental disorders and intellectual disability where extensive heterogeneity is observed [Soden et al., 2014; Gilissen et al., 2014]. In addition, the WGS methods can shed light on to problematic genomic regions that is currently not captured by gene panels or other conventional approaches. For example, Neonatal Intensive Care Unit (NICU) currently uses karyotyping, chromosomal microarrays, gene panels and single-gene tests to identify genetic abnormalities in newborns. The genome-wide sequencing is used in only a small fraction of newborns in NICU and in many cases, both conventional approaches and gene panels have limited diagnostic utility and warrants a WGS approach [Berg et al., 2017]. STATseq is a WGS method that can rapidly sequence patient samples within 50 h and has been shown to provide definitive diagnoses in 57% of the 35 cases on genomic regions that were not captured by standard genetic tests [Willig et al., 2015]. While genome-scale sequencing methods can provide comprehensive genetic information specific to each individual, interpretation of the data becomes difficulty and can lead to ambiguity especially if there is lack of uniformly accepted guidelines [Khotskaya et al., 2017].

Applications of Long-read technologies

While the NGS technology can generate vast amount of high quality sequence data at a lower sequencing cost, the shorter length of reads generated by these technologies are difficult to use in applications such as *de novo* genome assembly, haplotyping, identifying large structural variations (SV), sequencing long repeats or identifying full-length isoforms where the spliced exons are present in a contiguous linear read. The presence of repetitive regions that are multi-kilobase in length leads to alignment problems resulting in incompletely assembled genome due to low sequence coverage. Complete assemblies can only be built if there is sufficient information that connects two ends of the scaffolds. This is one of the main drawbacks of short-read technologies but can be complemented by the longer reads offered by both PacBio and ONT technologies. The long-read technologies can either be used alone or in combination with short-reads in a hybrid approach to resolve longer genomic structural features [Goodwin et al., 2015; Koren et al., 2012].

The shorter read lengths offered by the NGS platforms make it better suited to study small structural variants such as single nucleotide polymorphisms (SNPs) or indels that are within the limits of these technologies but, large variations such as complex genomic rearrangements as often observed in the cancer genomes are quite challenging to study. Understanding these genomic events has huge implications for human health and in order them, the sequence reads must be long and capable of providing information over very long distances [Harel & Lupski, 2018]. The mate-pair sequencing using short reads has been used to address this problem that links two ends of long DNA fragments and brings them in close proximity. While this method captures some SVs over long distances, challenges still remain when assessing ultralong SVs and repeat regions [Korbel et al., 2007]. Long-read technologies have been successfully applied to resolve and identify novel structural variants, close gaps in the human genome and improve haplotyping [Chaisson et al., 2015; Jain et al., 2018; Merker et al., 2018].

The NGS technologies have served as the gold standard to characterize the transcriptome but both PacBio and ONT are currently emerging in the field of transcriptomics. The nanopore sequencing was used by Bolisetty *et al.,* [2015] to sequence cDNAs from four genes that codes for least to most complex alternative splice isoforms in the *Drosophila melanogaster* genome. The SMRT sequencing was used to characterize alternative splice isoforms of neurexins in greater detail [Treutlein et al., 2014]. In addition, these long-read technologies can be applied to detect modifications present in native DNA and RNA. The PCR amplification used to enrich DNA library by current NGS methods erases methylation marks but ONT has been shown to identify 5-methylcytosines and is based on the changes in ionic current between modified and unmodified bases [Jain et al., 2018]. The PacBio technology uses inter-pulse duration, the time taken by polymerase to incorporate two successive nucleotides during sequencing, to distinguish hydroxymethyl cytosine, 5-methyl cytosine from cytosines and methylated adenosines from adenosines [Flusberg et al., 2010]. The single molecule long-read technologies are also capable of sequencing direct RNA molecules and detect RNA modifications. For example, the HIV reverse transcriptase was used instead of DNA polymerase to sequence RNA molecules and modifications in PacBio [Vilfan et al., 2013] but ONT uses the nanopores, similar to DNA sequencing, to sequence RNA molecules [Garalde et al., 2018].

Conclusions

The field of sequencing has seen a tremendous growth over the past two decades that led to a constant influx of new technologies. With rapid and mass adoption of these technologies, researchers have been able to address many clinically important and biologically relevant questions and advanced the quest for science. The next-generation sequencing technologies have clearly revolutionized the field of genomics and the first-generation technologies have been the cornerstone for these technologies that led to a dramatic increase in data throughput and rapid decline in sequencing costs. The wave of third-generation technologies took off recently,

but while these platforms currently suffer from high cost per base, low quality and throughput compared to other NGS technologies, TGS platforms clearly have the potential to advance genomics. Rapid technological advancements associated with TGS platforms accompanied by lower sequencing costs can open up new avenues for research such as measuring the abundance of transcripts at individual isoform level, studying exon connectivity across distant alternative exons, and the modifications on DNA and RNA on native molecules which are currently not possible on the NGS platforms.

CHAPTER 2

Determining exon connectivity in complex mRNAs by nanopore sequencing

Bolisetty, Mohan T., Gopinath Rajadinakaran, and Brenton R. Graveley. Determining Exon Connectivity in Complex mRNAs by Nanopore Sequencing. *Genome Biology* 16:204 (2015)

Abstract

Short-read high-throughput RNA sequencing, though powerful, is limited in its ability to directly measure exon connectivity in mRNAs that contain multiple alternative exons located farther apart than the maximum read length. Here, we use the Oxford Nanopore MinION sequencer to identify 7,899 'full-length' isoforms expressed from four Drosophila genes, *Dscam1, MRP, Mhc, and Rdl*. These results demonstrate that nanopore sequencing can be used to deconvolute individual isoforms and that it has the potential to be a powerful method for comprehensive transcriptome characterization.

Background

High throughput RNA sequencing has revolutionized genomics and our understanding of the transcriptomes of many organisms. Most eukaryotic genes encode pre-mRNAs that are alternatively spliced [Nilsen & Graveley, 2010]. In many genes, alternative splicing occurs at multiple places in the transcribed pre-mRNAs that are often located farther apart than the read lengths of most current high throughput sequencing platforms. As a result, several transcript assembly and quantitation software tools have been developed to address this [Trapnell et al.,

2010; Grabherr et al., 2011]. While these computational approaches do well with many transcripts, they generally have difficulty assembling transcripts of genes that express many isoforms. In fact, we have been unable to successfully assemble transcripts of complex alternatively spliced genes such as *Dscam1* or *Mhc* using any transcript assembly software (data not shown). These software tools also have difficulty quantitating transcripts that have many isoforms, and for genes with distantly located alternatively spliced regions, they can only infer, and not directly measure, which isoforms may have been present in the original RNA sample [Garber et al., 2011]. For example, consider a gene containing two alternatively spliced exons located 2 kbp away from one another in the mRNA. If each exon is observed to be included at a frequency of 50 % from short read sequence data, it is impossible to determine whether there are two equally abundant isoforms that each contain or lack both exons, or four equally abundant isoforms that contain both, neither, or only one or the other exon.

Pacific Bioscience sequencing can generate read lengths sufficient to sequence full length cDNA isoforms and several groups have recently reported the use of this approach to characterize the transcriptome [Sharon et al., 2013]. However, the large capital expense of this platform can be a prohibitive barrier for some users. Thus, it remains difficult to accurately and directly determine the connectivity of exons within the same transcript. The MinION nanopore sequencer from Oxford Nanopore requires a small initial financial investment, can generate extremely long reads, and has the potential to revolutionize transcriptome characterization, as well as other areas of genomics.

Several eukaryotic genes can encode hundreds to thousands of isoforms. For example, in *Drosophila*, 47 genes encode over 1,000 isoforms each [Brown et al., 2014]. Of these, *Dscam1* is the most extensively alternatively spliced gene known and contains 115 exons, 95 of which are alternatively spliced and organized into four clusters [Schmucker et al., 2000]. The exon 4, 6, 9, and 17 clusters contain 12, 48, 33, and 2 exons, respectively. The exons within each cluster are

spliced in a mutually exclusive manner and *Dscam1* therefore has the potential to generate 38,016 different mRNA and protein isoforms. The variable exon clusters are also located far from one another in the mRNA and the exons within each cluster are up to 80 % identical to one another at the nucleotide level. Together, these characteristics present numerous challenges to characterize exon connectivity within full-length *Dscam1* transcripts for any sequencing platform. Furthermore, though no other gene is as complex as *Dscam1*, many other genes have similar issues that confound the determination of exon connectivity.

We are interested in developing methods to perform simple and robust long-read sequencing of individual isoforms of *Dscam1* and other complex alternatively spliced genes. Here, we use the Oxford Nanopore MinION to sequence 'full-length' cDNAs from four *Drosophila* genes – *Rdl, MRP, Mhc*, and *Dscam1* – and identify a total of 7,899 distinct isoforms expressed by these four genes.

Results and discussion

Similarity between alternative exons We were interested in determining the feasibility of using the MinION nanopore sequencer to characterize the connectivity of distantly located exons in the mRNAs expressed from genes with complex splicing patterns. For the purposes of these experiments, we have focused on four *Drosophila* genes with increasingly complex patterns of alternative splicing (Fig. 2.1). Resistant to dieldrin (*Rdl*) contains two clusters, each containing two mutually exclusive exons and therefore has the potential to generate four different isoforms (Fig. 2.1a). Multidrug-Resistance like Protein 1 (*MRP*) contains two mutually exclusive exons in cluster 1 and eight mutually exclusive exons in cluster 2, and can generate 16 possible isoforms (Fig. 2.1b). Myosin heavy chain (*Mhc*) can potentially generate 180 isoforms due to five clusters of mutually exclusive exons – clusters 1 and 5 contain two exons, clusters 2 and 3 each contain

three exons, and cluster 4 contains five exons. Finally, *Dscam1* contains 12 exon 4 variants, 48 exon 6 variants, 33 exon 9 variants (Fig. 2.1d), and two exon 17 variants (not shown) and can potentially express 38,016 isoforms. For this study, however, we have focused only on the exon 3 through exon 10 region of *Dscam1*, which encompasses the 93 exon 4, 6, and 9 variants, and 19,008 potential isoforms (Fig. 2.1d).

Because our nanopore sequence analysis pipeline uses LAST to perform alignments [Frith et al., 2010], we aligned all of the *Rdl, MRP, Mhc*, and *Dscam1* exons within each cluster to one another using LAST to determine the extent of discrimination needed to accurately assign nanopore reads to a specific exon variant. For *Rdl*, each variable exon was only aligned to itself, and not to the other exon in the same cluster (data not shown). For *MRP*, the two exons within cluster 1 only align to themselves, and though the eight variable exons in cluster 2 do align to other exons, there is sufficient specificity to accurately assign nanopore reads to individual exons (Fig. 2.2a). For *Mhc*, the variable exons in cluster 1 and cluster 5 do not align to other exons, and the variable exons in cluster 2, cluster 3, and cluster 4 again align with sufficient discrimination to identify the precise exon present in the nanopore reads (Fig. 2.2b). Finally, for *Dscam1*, the difference in the LAST alignment scores between the best alignment (each exon to itself) and the second, third, and fourth best alignments are sufficient to identify the *Dscam1* exon variant (Fig. 2.2c). This analysis indicates that for each gene in this study, LAST alignment scores are sufficiently distinct to identify the variable exons present in each nanopore read.

Optimizing template switching in *Dscam1* cDNA libraries

Template switching can occur frequently when libraries are prepared by PCR and can confound the interpretation of results [McManus et al., 2010; Plocik et al., 2013]. For example,

Figure 2.1 Schematic of the exon-intron structures of the genes examined in this study. **a** The *Rdl* gene contains two clusters (cluster one and two) which each contain two mutually exclusive exons. **b** The *MRP* gene contains contains two and eight mutually exclusive exons in clusters 1 and 2, respectively. **c** *Mhc* contains two mutually exclusive exons in clusters 1 and 5, three mutually exclusive exons in clusters 2 and 3, and five mutually exclusive exons in cluster 4. **d** The *Dscam1* gene contains 12, 48, and 33 mutually exclusive exons in the exon 4, 6, and 9 clusters, respectively. For each gene, the constitutive exons are colored blue, while the variable exons are colored yellow, red, orange, green, or light blue

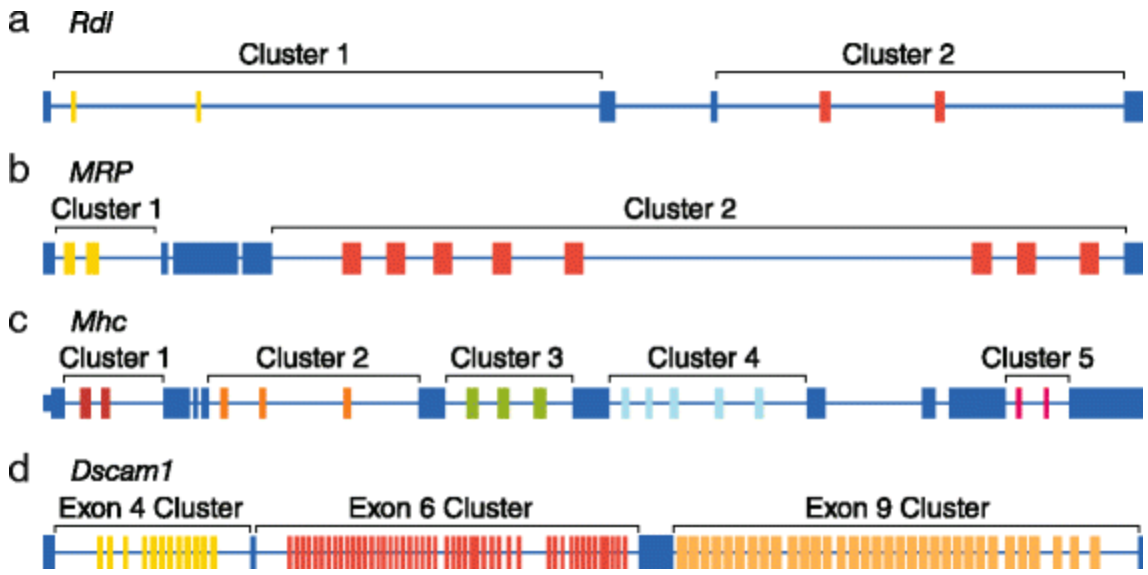
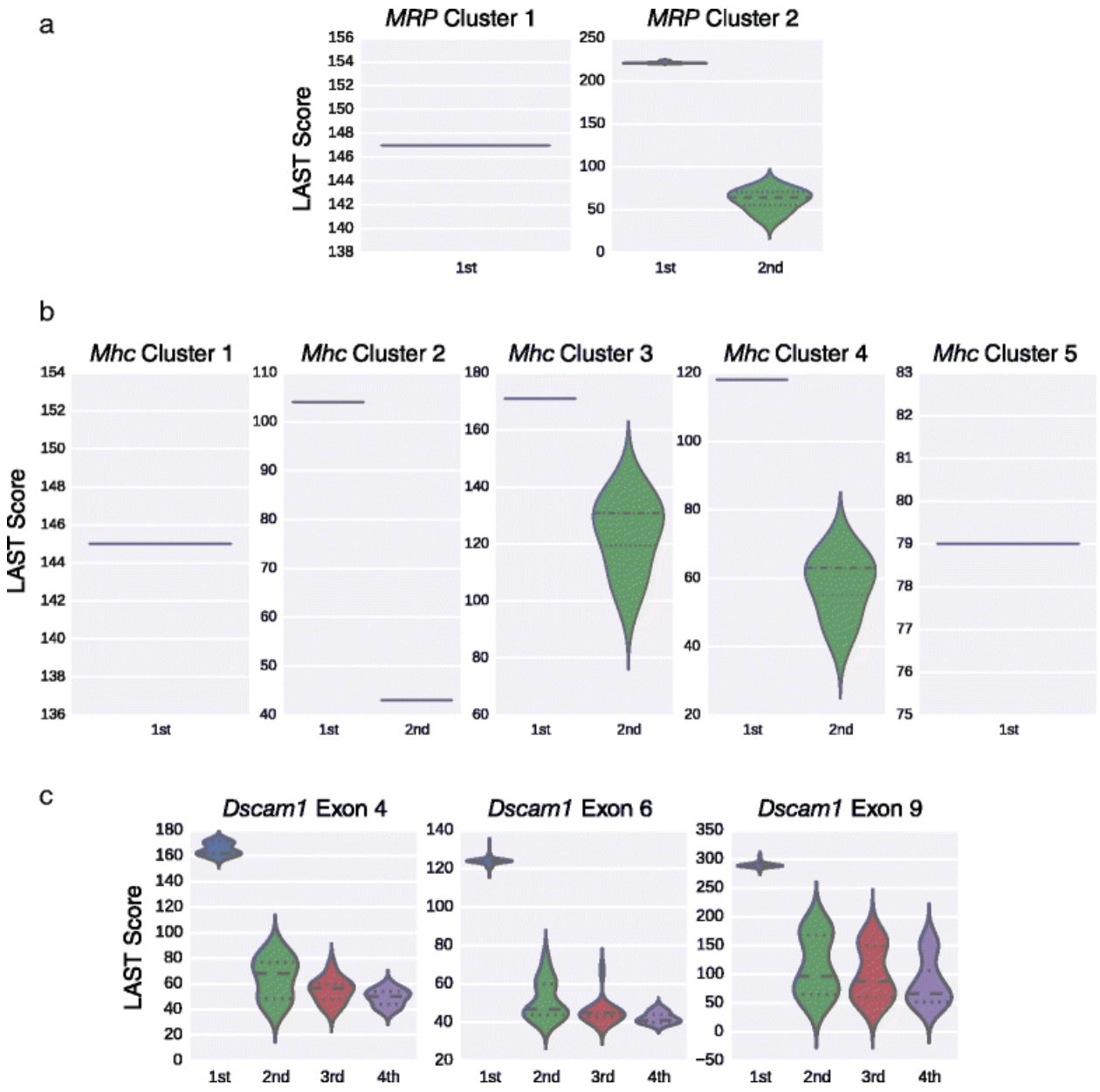

a
Rdl
Cluster 1
Cluster 2
b
MRP
Cluster 1
Cluster 2
c
Mhc
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
d
Dscam1
Exon 4 Cluster
Exon 6 Cluster
Exon 9 Cluster

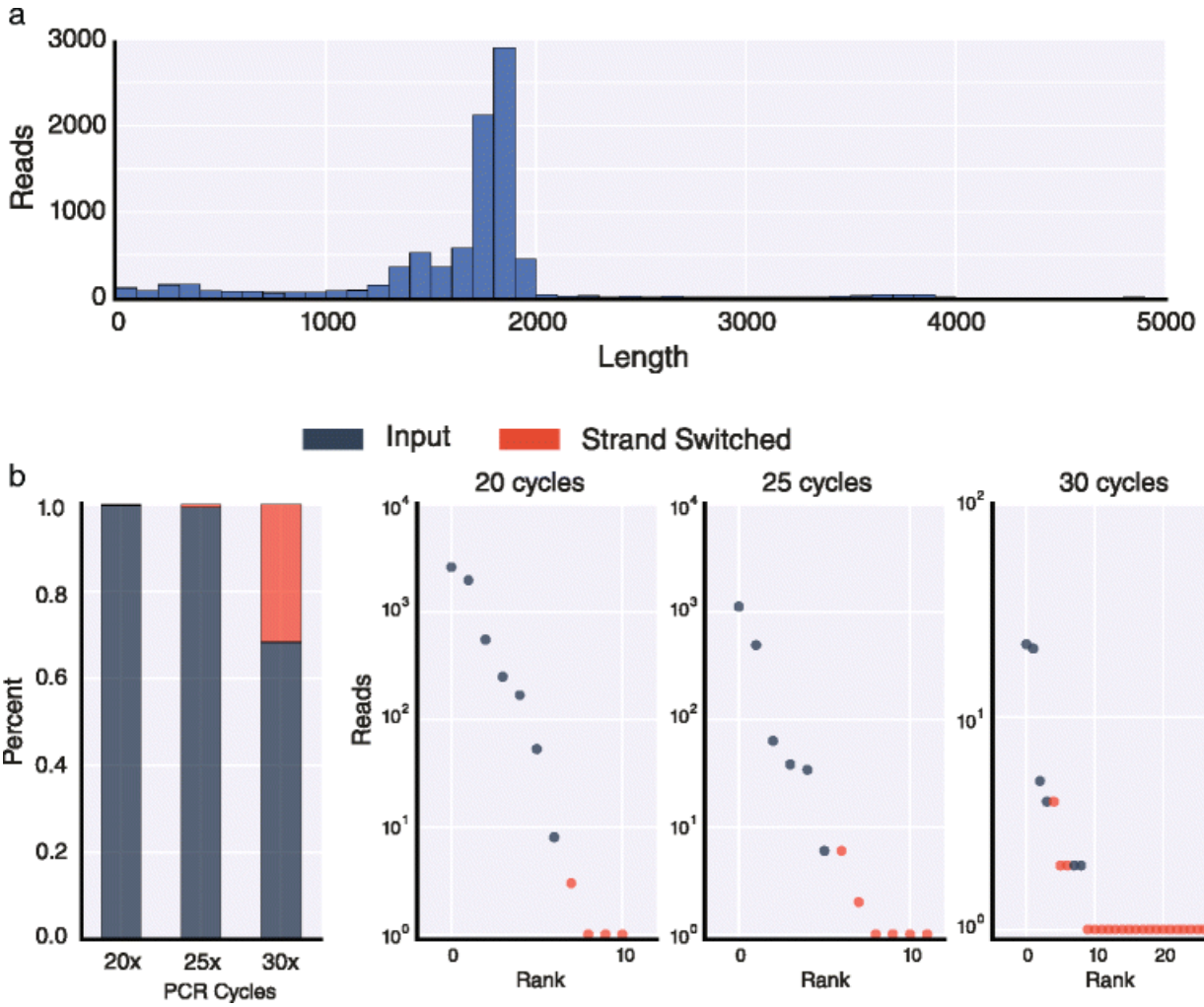
Figure 2.2 Similarity distance between the variable alternative exons of *MRP*, *Mhc*, and *Dscam1*. **a** Violin plots of the LAST alignment scores of each variable exon within *MRP* cluster 1 and MRP cluster 2 to themselves and the second (2nd) best alignments. **b** Violin plots of the LAST alignment scores of each variable exon within each *Mhc* cluster to themselves and the second (2nd) best alignments. **c** Violin plots of the LAST alignment scores of each variable exon within each *Dscam1* cluster to themselves (1st), and to the exons with the second (2nd), third (3rd) and fourth (4th) best alignments


a
MRP Cluster 1
MRP Cluster 2
LAST Score
1st
1st
2nd
b
Mhc Cluster 1
Mhc Cluster 2
Mhc Cluster 3
Mhc Cluster 4
Mhc Cluster 5
LAST Score
c
Dscam1 Exon 4
Dscam1 Exon 6
Dscam1 Exon 9
LAST Score
1st
2nd
3rd
4th

CAM-Seq [Sun et al., 2013] and a similar method we independently developed called TripleRead sequencing [Roy et al., 2015] to characterize *Dscam1* isoforms, were found to have excessive template switching due to amplification during the library prep protocols. To assess template switching in our current study, we generated a spike-in mixture of in vitro transcribed RNAs representing six unique *Dscam1* isoforms – *Dscam1*$^{4.2,6.32,9.31}$, *Dscam1*$^{4.1,6.46,9.30}$, *Dscam1*$^{4.3,6.33,9.9}$, *Dscam1*$^{4.12,6.44,9.32}$, *Dscam1*$^{4.7,6.8,9.15}$, and *Dscam1*$^{4.5,6.4,9.4}$. We used 10 pg of this control spike-in mixture and prepared libraries for MinION sequencing by amplifying the exon 3 through exon 10 region for 20, 25, or 30 cycles of RT-PCR. We then end-repaired and dA-tailed the fragments, ligated adapters, and sequenced the samples on a MinION (7.3) for 12 h each. We obtained 33,736, 8,961, and 7,511 basecalled reads from the 20, 25, and 30 cycle libraries, respectively. Consistent with the size of the exon 3 to 10 cDNA fragment being 1,806–1,860 bp in length, depending on the precise combination of exons it contains, most reads we observed were in this size range (Fig. 2.3a). We used Poretools [Loman & Quinlan, 2014] to convert the raw output files into fasta format and then used LAST to align the reads to a LAST database containing each variable exon. From these alignments, we identified reads that mapped to all three exon clusters, as well as the exon with the best alignment score within each cluster. When examining the alignments to each cluster independently, we found that for these spike-in libraries, all reads mapped uniquely to the exons present in the input isoforms. Therefore, any observed isoforms that were not present in the input pool were a result of template switching during the RT-PCR and library prep protocol and not due to false alignments or sequencing errors.

When comparing the combinations of exons within each read to the input isoforms, we observed that 32 % of the reads from the 30 cycle library corresponded to isoforms generated by template switching (Fig. 2.3b). The template-switched isoforms observed by the greatest number of reads in the 30 cycle library were due to template switching between the two most frequently sequenced input isoforms. In most cases, template switching occurred somewhere within exon 7

Figure 2.3 Optimized RT-PCR minimizes template-switching for MinION sequencing. **a** Histogram of read lengths from MinION sequencing of *Dscam1* spike-ins from the library generated using 25 cycles of PCR. **b** Bar plot indicating the extent of template switching in *Dscam1* spike-ins at different PCR cycles (left). The blue portions indicate the fraction of reads corresponding to input isoforms while the red portions correspond to the fraction of reads corresponding to template-switched isoforms. On the right, plots of the rank order versus number of reads (log10) for the 20, 25, and 30 cycle libraries. The blue dots indicate input isoforms while the red portions correspond to template-switched isoforms


a
Reads
3000
2000
1000
0
0
1000
2000
3000
4000
5000
Length
Input
Strand Switched
b
Percent
1.0
0.8
0.6
0.4
0.2
0.0
20x
25x
30x
PCR Cycles
20 cycles
25 cycles
30 cycles
Reads
Rank
Rank
Rank

or 8 and resulted in a change in exon 9. However, the extent of template switching was reduced to only 1 % in the libraries prepared using 25 cycles, and to 0.2 % in the libraries prepared using 20 cycles of PCR (Fig. 2.3b). Again, for these two libraries the most frequently sequenced template-switched isoforms involved the input isoforms that were also the most frequently sequenced. These experiments demonstrate that the MinION nanopore sequencer can be used to sequence 'full length' Dscam1 cDNAs with sufficient accuracy to identify isoforms and that the cDNA libraries can be prepared in a manner that results in a very small amount of template switching.

Dscam1 isoforms observed in adult heads

To explore the diversity of *Dscam1* isoforms expressed in a biological sample, we prepared a *Dscam1* library from RNA isolated from D. melanogaster heads prepared from mixed male and female adults using 25 cycles of PCR and sequenced it for 12 h on the MinION nanopore sequencer obtaining a total of 159,948 reads of which 78,097 were template reads, 48,474 were complement reads, and 33,377 were 2D reads (Fig. 2.4a). We aligned the reads individually to the exon 4, 6, and 9 variants using LAST. A total of 28,971 reads could be uniquely or preferentially aligned to a single variant in all three clusters. For further analysis, we used all 16,419 2D read alignments and 31 1D reads when both template and complement aligned to same variant exons (not all reads with both a template and complement yield a 2D read). The remaining 12,521 aligned reads were 1D reads where there was either only a template or complement read, or when the template and complement reads disagreed with one another and were therefore not used further. We observed 92 of the 93 potential exon 4, 6, or 9 variants – only exon 6.11 was not observed in any read (Fig. 2.4f). To assess the accuracy of the results we performed RT-PCR using primers in the flanking constitutive exons that contained Illumina sequencing primers to separately amplify the Dscam1 exon 4, 6, and 9 clusters from the same

RNA used to prepare the MinION libraries, and sequenced the amplicons on an Illumina MiSeq. The frequency of variable exon use in each cluster was extremely consistent between the two methods (R^2 = 0.95, Fig. 2.5a).

Over their entire lengths, the 2D reads that map specifically to one exon 4, 6, and 9 variants map with an average 90.37 % identity and an average LAST score of approximately 1,200 (Fig. 2.5b). The 16,450 full length reads correspond to 7,874 unique isoforms, or 42 % of the 18,612 possible isoforms given the exon 4, 6, and 9 variants observed. We note, however, that while 4,385 isoforms were represented by more than one read, 3,516 of isoforms were represented by only one read indicating that the depth of sequencing has not reached saturation (Fig. 2.4b and 2.4c). This was further confirmed by performing a bootstrapped subsampling analysis (Fig. 2.4d) and by using the capture-recapture method to attempt to assess the complexity of isoforms present in the library (Fig. 2.4e), which suggests that over 11,000 isoforms are likely to be present, though even this analysis has not yet reached saturation. The most frequently observed isoforms were *Dscam1*$^{4.1,6.12,9.30}$ and *Dscam1*$^{4.1,6.1,9.30}$ which were observed with 30 and 25 reads, respectively (Fig. 2.4e). In conclusion, these results demonstrate the practical application of using the MinION nanopore sequencer to identify thousands of distinct *Dscam1* isoforms in a single biological sample.

Nanopore sequencing of 'full-length' *Rdl, MRP*, and *Mhc* isoforms

To extend this approach to other genes with complex splicing patterns, we focused on *Rdl, MRP*, and *Mhc* which have the potential to generate four, 16, and 180 isoforms, respectively. We prepared libraries for each of these genes by RT-PCR using primers in the constitutive exons flanking the most distal alternative exons using 25 cycles of PCR, pooled the three libraries and sequenced them together on the MinION nanopore sequencer for 12 h obtaining a total of 22,962

Figure 2.4 MinION sequencing of *Dscam1* identified 7,874 isoforms. **a** Histogram of read length distribution for *Drosophila* head samples. **b** The total number of *Dscam1* isoforms identified from MinION sequencing. **c** Cumulative distribution of *Dscam1* isoforms with respect to expression. **d** Violin plot of the number of isoforms identified using 100 random pools of the indicated number of reads. **e** Plot of the estimated number of total isoforms present in the library using the capture-recapture method with two random pools of the indicated number of reads. The shaded blue area indicates the 95 % confidence interval. **f** Deconvoluted expression of *Dscam1* exon cluster variants (top) and the isoform connectivity of two highly expressed *Dscam1* isoforms (bottom)

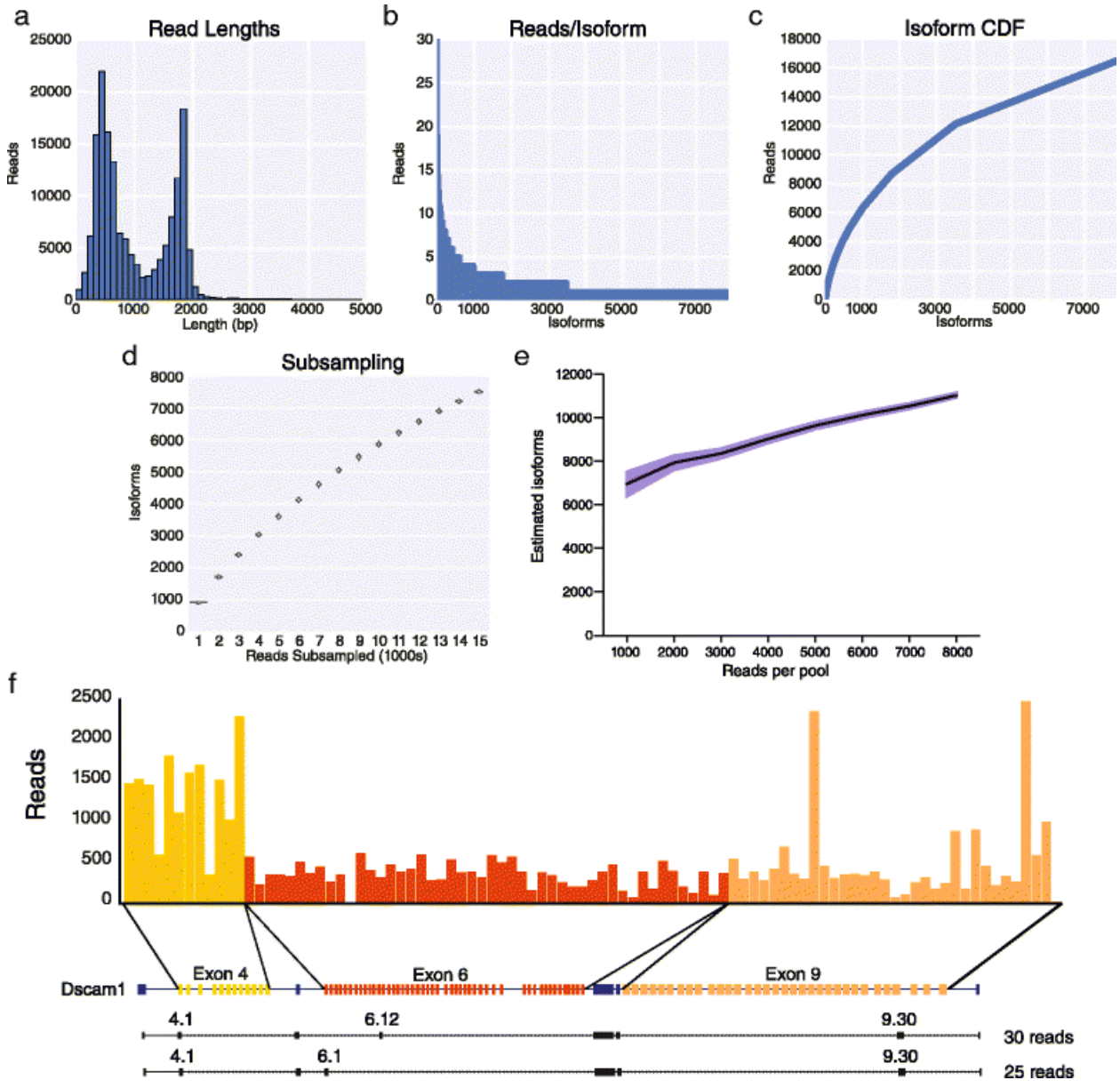
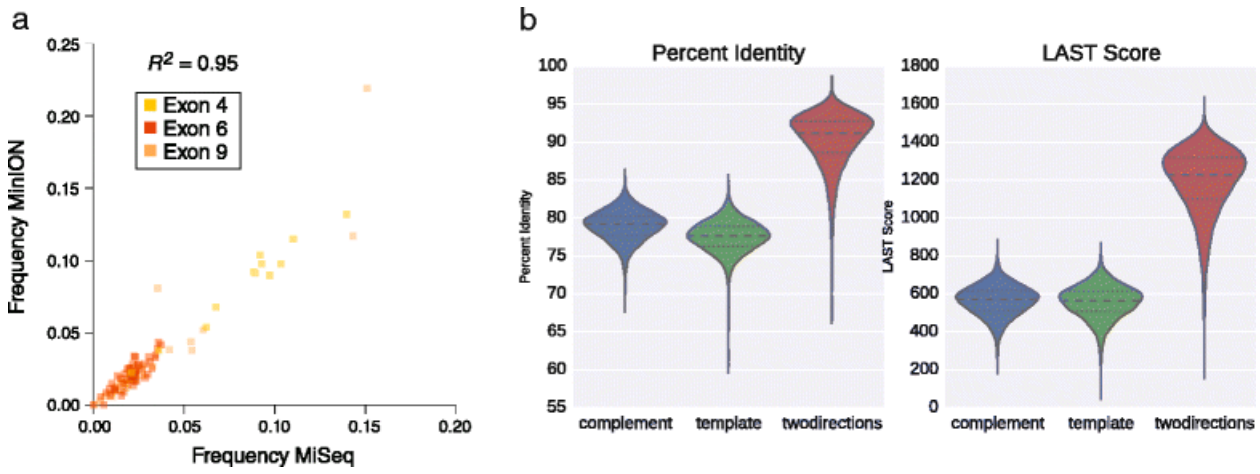

a
Read Lengths
Reads
Length (bp)
b
Reads/Isoform
Reads
Isoforms
c
Isoform CDF
Reads
Isoforms
d
Subsampling
Isoforms
Reads Subsampled (1000s)
e
Estimated Isoforms
Reads per pool
f
Reads
Exon 4
Exon 6
Exon 9
Dscam1
4.1
6.12
9.30
30 reads
4.1
6.1
9.30
25 reads

Figure 2.5 Accuracy of *Dscam1* sequencing results. **a** Comparison of the frequency of variable exon inclusion for the *Dscam1* exon 4 (yellow), 6 (red), and 9 (orange) clusters as determined by nanopore sequencing or by amplicon sequencing using an Illumina MiSeq. **b** Percent identities (left) or LAST alignment scores (right) of full-length template, complement, and two directions (sequencing both template and complements) nanopore read alignments


a
$R^2 = 0.95$
Exon 4
Exon 6
Exon 9
Frequency MinION
Frequency MiSeq
0.00
0.05
0.10
0.15
0.20
0.25
b
Percent Identity
Percent Identity
LAST Score
LAST Score
complement
template
twodirections

reads. The input libraries for *Rdl, MRP*, and *Mhc* were 567 bp, 1,769-1,772 bp, and 3,824 bp, respectively. The raw reads were aligned independently to LAST indexes of each cluster of variable exons. The alignment results were then used to assign reads to their respective libraries, identify reads that mapped to all variable exon clusters for each gene, and the exon with the best alignment score within each cluster. In total, we obtained 301, 337, and 112 full length reads for *Rdl* (Fig. 2.6), *MRP* (Fig. 2.7), and *Mhc* (Fig. 2.8), respectively. For *Rdl*, both variable exons in each cluster was observed, and accordingly all four possible isoforms were observed, though in each case the first exon was observed at a much higher frequency than the second exon (Fig. 6d). Interestingly, the ratio of isoforms containing the first versus second exon in the second cluster is similar for isoforms containing either the first exon or the second exon in the first cluster indicating that the splicing of these two clusters may be independent. For *MRP*, both exons in the first cluster were observed and all but one of the exons in the second cluster (exon B) were observed, though the frequency at which the exons in both clusters were used varied dramatically (Fig. 2.7d). For example, within the first cluster, exon B was observed 333 times while exon A was observed only four times. Similarly, in the second cluster, exon A was observed 157 times whereas exons B, E, F, and G were observed 0 times, thrice, once, and twice, respectively, and exons D, E, and H were observed between 40 and 76 times. As a result, we observed only nine MRP isoforms. For *Mhc*, we again observed strong biases in the exons observed in each of the five clusters (Fig. 2.8d). In the first cluster, exon B was observed more frequently than exon A. In the second cluster, 109 of the reads corresponded to exon A, while exons B and C were observed by only two and one read, respectively. In the third cluster, exon A was not observed at all while exons B and C were observed in roughly 80 % and 20 % of reads, respectively. In the fourth cluster, exon A was observed only once, exons B and C were not observed at all, exon E was observed 13 times while exon D was present in all of the remaining reads. Finally, in the fifth cluster, only exon B was observed. As with *MRP*, these strong biases and near or complete absences of exons in some of the clusters severely reduces the number of possible isoforms that

Figure 2.6 MinION sequencing of *Rdl* identified four isoforms. **a** Histogram of read lengths. **b** The number of reads per isoform. **c** Cumulative distribution of isoforms with respect to expression. **d** The number of reads per alternative exon (top) and per isoform (below)

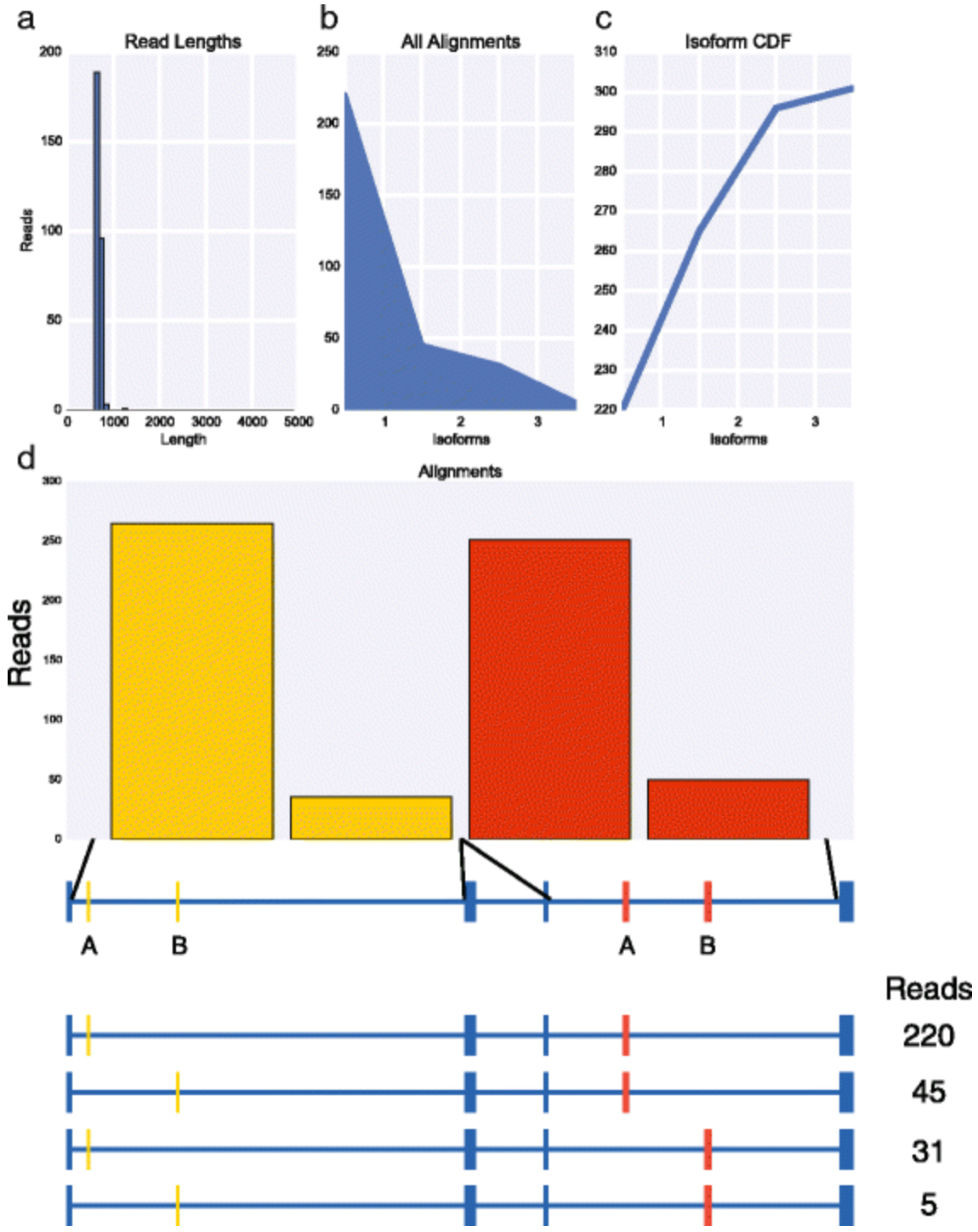

a
Read Lengths
Reads
Length
b
All Alignments
Isoforms
c
Isoform CDF
Isoforms
d
Alignments
Reads
A
B
A
B
Reads
220
45
31
5

Figure 2.7 MinION sequencing of *MRP* identified nine isoforms. **a** Histogram of read lengths. **b** The number of reads per isoform. **c** Cumulative distribution of isoforms with respect to expression. **d** The number of reads per alternative exon (top) and per isoform (below)

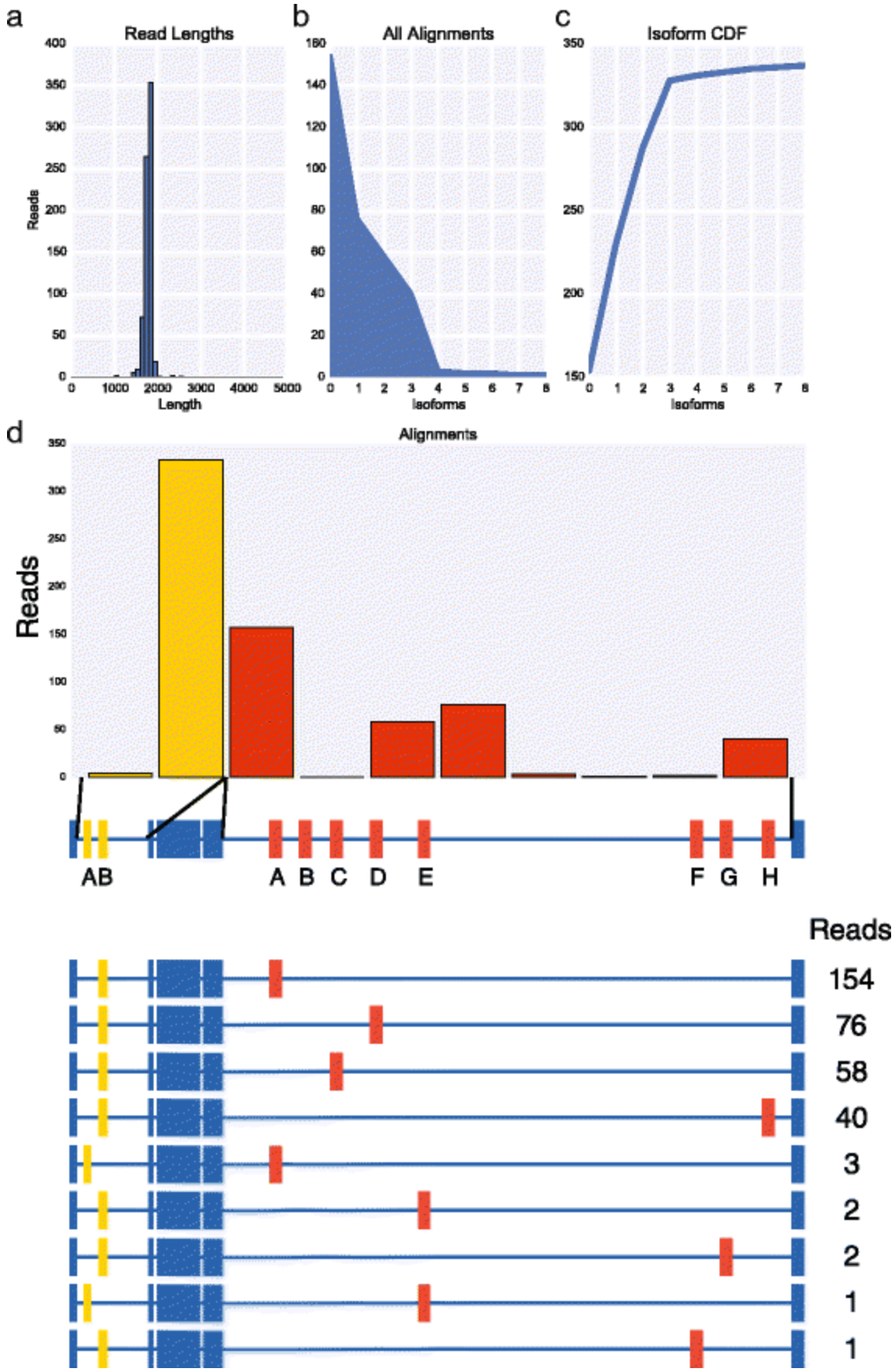
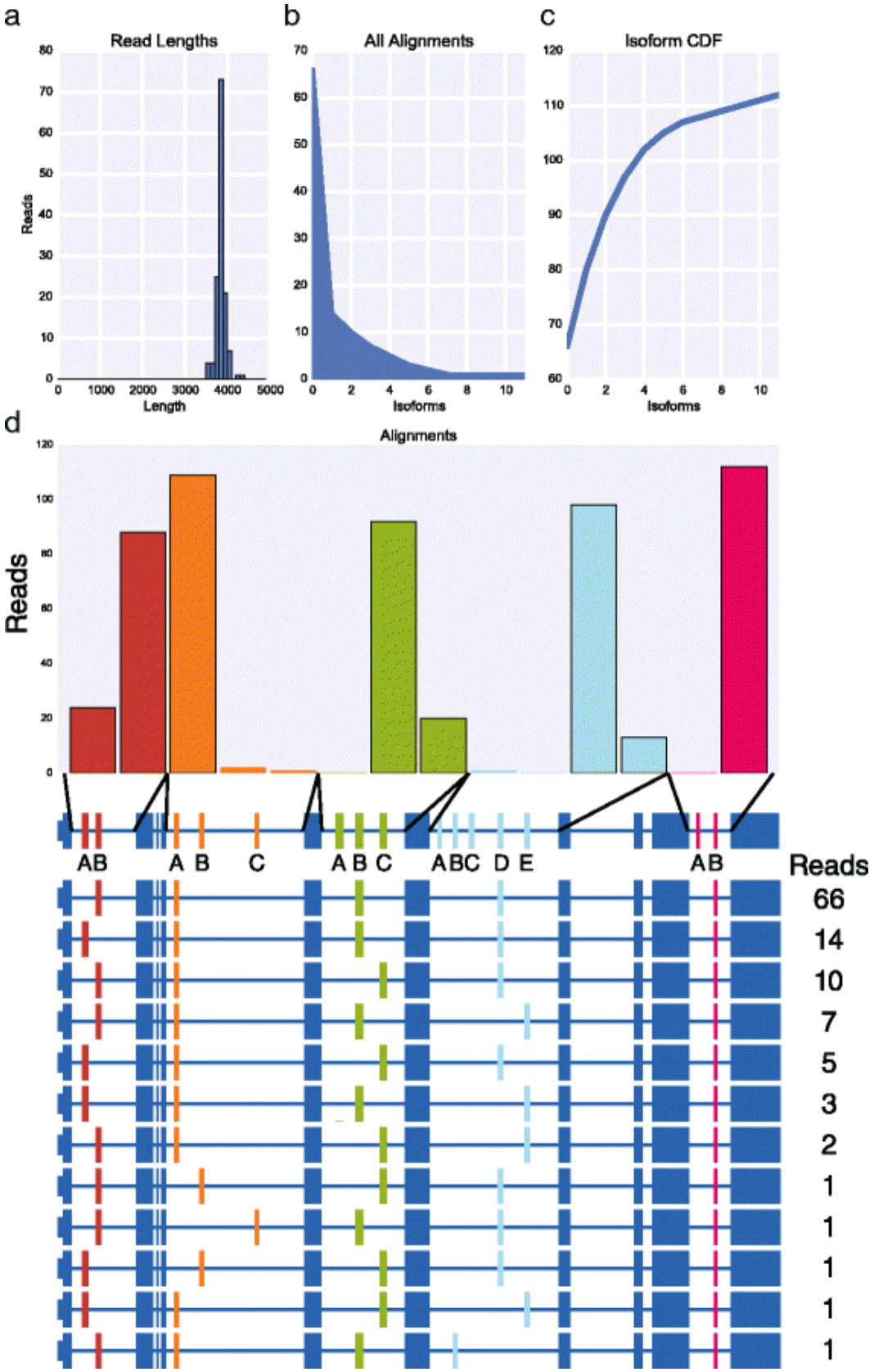

a
Read Lengths
Reads
Length
b
All Alignments
Isoforms
c
Isoform CDF
Isoforms
d
Alignments
Reads
AB
A B C D E
F G H
Reads
154
76
58
40
3
2
2
1
1

Figure 2.8 MinION sequencing of *Mhc* identified 12 isoforms. **a** Histogram of read lengths. **b** The number of reads per isoform. **c** Cumulative distribution of isoforms with respect to expression. **d** The number of reads per alternative exon (top) and per isoform (below)


a
Read Lengths
Reads
Length
b
All Alignments
Isoforms
c
Isoform CDF
Isoforms
d
Alignments
Reads
AB
A B
C
A B C
ABC
D E
AB
Reads
66
14
10
7
5
3
2
1
1
1
1
1

can be observed. In fact, of the 180 potential isoforms encoded by *Mhc*, we observed only 12 isoforms. Various *Mhc* isoforms are known to be expressed in striking spatial and temporally restricted patterns [Zhang et al., 2001] and thus it is likely that other *Mhc* isoforms that we did not observe, could be observed by sequencing other tissue samples.

Conclusions

Here we have demonstrated that nanopore sequencing with the Oxford Nanopore MinION can be used to easily determine the connectivity of exons in a single transcript, including *Dscam1*, the most complicated alternatively spliced gene known in nature. This is an important advance for several reasons. First, because short-read sequence data cannot be used to conclusively determine which exons are present in the same RNA molecule, especially for complex alternatively spliced genes, long-read sequence data are necessary to fully characterize the transcript structure and exon connectivity of eukaryotic transcriptomes. Second, although the Pacific Bioscience platform can perform long-read sequencing, there are several differences between it and the Oxford Nanopore MinION that could cause users to choose one platform over the other. In general, the quality of the sequence generated by the Pacific Bioscience is higher than that currently generated by the Oxford Nanopore MinION. This is largely due to the fact that each molecule is sequenced multiple times on the Pacific Bioscience platform yielding a high-quality consensus sequence whereas on the Oxford Nanopore MinION, each molecule is sequenced at most twice (in the template and complement). We have previously used the Pacific Bioscience platform to characterize Dscam1 isoforms and found that it works well, though due to the large amount of cDNA needed to generate the libraries, many cycles of PCR are necessary and we observed an extensive amount of template switching, making it impractical to use for these experiments (BRG, unpublished data). However, over the past year that we have been involved in the MAP, the quality of sequence has steadily increased. As this trend is likely to

continue, the difference in sequence quality between these two platforms is almost certain to shrink. Nonetheless, as we demonstrate, the current quality of the data is more than sufficient to allow us to accurately distinguish between highly similar alternatively spliced isoforms of the most complex gene in nature. Third, the ability to accurately characterize alternatively spliced transcripts with the Oxford Nanopore MinION makes this technology accessible to a much broader range of researchers than was previously possible. This is in part due to the fact that, in contrast to all other sequencing platforms, very little capital expense is needed to acquire the sequencer. Moreover, the MinION is truly a portable sequencer that could literally be used in the field (provided one has access to an Internet connection), and due to its size, almost no laboratory space is required for its use.

Although nanopore sequencing has many exciting and potentially disruptive advantages, there are several areas in which improvement is needed. First, although we were able to accurately identify over 7,000 *Dscam1* isoforms with an average identity of full-length alignments >90 %, there are several situations in which this level of accuracy will be insufficient to determine transcript structure. For instance, there are many micro-exons in the human genome [Irimia et al., 2014], and these exons would be difficult to identify if they overlapped a portion of a read that contained errors. Additionally, small unannotated exons could be difficult to identify for similar reasons. Second, the current number of usable reads is lower than that which will be required to perform whole transcriptome analysis. One issue that plagues transcriptome studies is that the majority of the sequence generated comes from the most abundant transcripts. Thus, with the current throughput, numerous runs would be needed to generate a sufficient number of reads necessary to sample transcripts expressed at a low level. In fact, this is one reason that we chose in this study, to begin by targeting specific genes rather than attempting to sequence the entire transcriptome. We do note, however, that over the past year of our participation in the MAP, the throughput of the Oxford Nanopore MinION has increased, and it is reasonable to expect

additional improvements in throughput that should make it possible to generate a sufficient number of long reads to deeply interrogate even the most complex transcriptome.

In conclusion, we anticipate that nanopore sequencing of whole transcriptomes, rather than targeted genes as we have performed here, will be a rapid and powerful approach for characterizing isoforms, especially with improvements in the throughput and accuracy of the technology, and the simplification and/or elimination of the time-consuming library preparations.

Materials and methods

Drosophila strains

Drosophila melanogaster y; cn b sp (stock: 2057, Bloomington) were maintained and raised at room temperature.

Spike-in preparation

Total RNA from about 30 heads was extracted using Trizol reagent. One microgram of total RNA was used to synthesize cDNA using random hexamers with SuperScript II (Invitrogen, Cat No: 18064) in a 20 µL reaction; 2 µl of cDNA reaction was used to amplify Dscam1 exons 4 through 9 using the primers exon 3 and exon 10 with LongAmp (New England Biolabs, Cat No: M0323) in a 50 µL reaction volume with the following PCR condition: initial denaturation at 94 °C for 30 s, denaturation at 94 °C for 15 s, annealing at 58 °C for 15 s, extension at 65 °C for 100 s (40X cycle), final extension at 65 °C for 10 min. The PCR amplicons were purified using MinElute PCR purification kit (Qiagen) and eluted in 20 µL ultrapure water. The eluted amplicons were then cloned into a vector with both T7 and SP6 dual promoters (Life Technologies, Cat No: K4600) and transformed into Top10 shot cells. A total of 96 colonies were sequenced to identify exon

variant sequences in individual clones. Six individual colonies containing a single, non-overlapping, unique exon variants were used to make spike-in RNAs. The vector containing the *Dscam1* insert and the T7, SP6 promoter sequences were amplified using M13F and M13R primers. The SP6 oriented clones were individually amplified using T7 overhang primers to facilitate in vitro transcription of all clones from T7 promoter using transcription kit. Following transcription, 1 µL RNA (1 µg/µL) of each of the six clones were mixed and a 10-fold serial dilution was made with concentration ranging from 100 ng/µL to 1 pg/µL. cDNA was synthesized using SuperScript II (Invitrogen, Cat No: 18064) and a 2.5 µL cDNA from 10 pg/µL reaction was used in the 25 µL Phusion PCR with the following conditions: initial denaturation at 95 °C for 30 s, denaturation at 95 °C for 10 s, annealing at 64.7 °C for 12 s, extension at 72 °C for 40 s (20X, 25X, and 30X cycles), final extension at 72 °C for 5 min, using primers CGGATCCATTATCTCCCGGGACG (Dscam1 exon 3) and CGGATCCCTGGGCGAAGGCC (Dscam1 exon 10 reverse).

Amplicon library preparation and Oxford Nanopore sequencing

The library preparation for amplicon sequencing was done using SQK-MAP003 following manufacturer's protocol (ONT). Briefly, a total of 850 ng (spike-in) and 1 µg (mixed heads) in 80 µL was end repaired using NEBNext End Repair Module (New England Biolabs, Cat No: E6050) and followed by dA tailing using NEBNext dA Tailing Module (New England Biolabs, Cat No: E6053). The dA tailed amplicons were then adapter ligated in a total of 100 µL reaction volume and incubated at room temperature for 10 min. This reaction mixture was then purified using Agencourt AMPure XP (Beckman Coulter Inc., cat. no. A63880) beads and washed and eluted in nanopore supplied reagents in 25 µL ultrapure water. This pre-sequencing mix was added with the fuel mix and EP buffer and loaded on the R7.3 flow cell and sequenced.

Nanopore data analysis

Poretools (version 0.3.0) [Loman & Quinlan, 2014] was used to extract fasta reads from Basecalled fast5 files. Exon cluster specific LAST indices were made using lastdb with default parameters. The reads were then aligned using lastal independently to these LAST indices using the following parameters: −s 2 -T 0 -Q 0 -a 1. Reads that aligned to all three clusters were parsed from all alignments and used for further processing. The top scoring alignment was used for reads that aligned to multiple variants. iPython notebooks containing all the analysis and code are available at github/mohanbolisetty/dscam_nanopore. MAF files from LAST alignments were converted to SAM or PSL formats using maf-convert.py.

***Dscam1* variable exon amplicon library preparation and Illumina sequencing**

For the *Dscam1* MiSeq amplicon library, cDNA was synthesized using 1 µL RNA (600 ng/µL) from mixed heads using SuperScript II (Invitrogen, Cat No: 18064) in a 20 µL reaction. A total of 2.5 µL of cDNA was used to individually amplify the exon 4, 6, and 9 clusters with Phusion (NEB Inc., catalog no. M0530L) using the following PCR protocol: 95 °C for 30 s followed by 30 cycles of 95 °C for 10 s, 59 °C for 12 s and 72 °C for 15 s, followed by a 5 min incubation at 72 °C using the following primer pairs:

Cluster4_Fwd:
AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACG
CTCTTCCGATCTATCggcaataccaggtactttcc

Cluster4_Rev:
CAAGCAGAAGACGGCATACGAGATCTAGTACGGTGACTGGAGTTCAGACGTGTGCTC
TTCCGATCTATCgatccattatctcccggga

Cluster6_Fwd:

AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGAC

GCTCTTCCGATCTATCtgttccttcgatgaacttgt

Cluster6_Rev:

CAAGCAGAAGACGGCATACGAGATCATGCCTAGTGACTGGAGTTCAGACGTGTGCTC

TTCCGATCTATCttaagtgccacaaaaggacg

Cluster9_Fwd:

AATGATACGGCGACCACCGAGATCTACACACCTCTTCACACTCTTTCCCTACACGACGCTC

TTCCG ATCTTCctcgaggatccatctggg

Cluster9_Rev:

CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTGACTGGAGTTCAGACGTGTGCTCTTC

CGATCT TCtcgaggatctctggaagtg

Following amplification, three separate PCR reactions were mixed together and purified using Agencourt AMPure XP (Beckman Coulter Inc., cat. no. A63880) beads. A library concentration of 2.1 nM was loaded and sequenced using MiSeq® Reagent Kit v3 (Illumina Inc., cat no. MS-102-3001).

MiSeq data analysis

The fastq files were processed in R using the package Biostrings [Pages et al.]. The reverse primer sequences from each of the Dscam1 exon 4, 6, and 9 clusters were matched (allowing no mismatches) against fasta sequences from read 2. The matching reads were subsequently aligned against each reference exon variant (length trimmed to 51 bp from the start of each variant) within a cluster for all three clusters.

Accession number

The raw nanopore data are available at the European Nucleotide Archive (ENA) under accession number ERP011508.

CHAPTER 3

Assessing the utility of Oxford Nanopore Platform for long-read DNA and direct RNA sequencing

Abstract

The Oxford Nanopore Technology (ONT), capable of generating longer reads enable the identification of exon connectivity over multiple distant transcript regions. To assess the performance of ONT platform for quantitative studies, we prepared replicate cDNA libraries from Spike-In RNA Variant (SIRV) mixes and sequenced them on the MinION. The SIRV mix contains 69 transcripts from 7 genes which are grouped into four sub-mixes that vary in their relative concentration over 8 orders of magnitude. The nanopore reads obtained from SIRV cDNA libraries were studied in detail and we were able to align over 94% of reads to SIRV transcriptome with tuned options using LAST. We were able to identify and manually assign between 17.5% and 26% of full-length 2D reads to their isoforms. The replicates correlated well with each other with a Pearson coefficient of 0.965, but the lower throughput and smaller fraction of full-length reads did not permit a comprehensive evaluation of relative quantity assessments between input RNA and the assigned reads. We additionally sequenced amplicon libraries generated from isoforms of 11 *Drosophila* ultracomplex genes (UCGs) and were able to identify previously unannotated exons. The long reads obtained from 4 UCGs also enabled identification of individual and combinatorial RNA editing sites in the *Ndae1* gene. By using direct RNA sequencing on yeast *Eno2*, we demonstrate that the ONT platform can be used to sequence RNA molecules directly without the need to synthesize cDNAs and that a vast majority of the high-quality RNA reads can be aligned full-length to the reference.

Introduction

Alternative splicing (AS) increases the coding potential of genes and can result in increased isoform and protein diversity [Nilsen & Graveley, 2010]. Transcriptome complexity is further increased by alternative promoter usage, alternative polyadenylation and RNA editing in addition to alternative splicing. The exons that are alternatively spliced can be located anywhere in the gene body. Co-regulated alternative splicing is a phenomenon by which alternative exons at different sites within a gene are spliced in a coordinated manner, leading to a unique combination of exons in the transcript (Fededa, 2005). Previous studies have suggested the role of co-regulated mode of splicing in human *Exoc7* (Fagnani, 2007) and mouse fibronectin (Fededa, 2005) but these studies were inconclusive because the use of microarray technology lacks the ability to detect exon connectivity and the mini-gene system used engineered gene constructs that lack complete regulatory control elements present *in vivo*. In another study using *C. elegans*, inclusion of alternative exons in *Slo1* at three different sites were found to be coordinated by assaying 12 *cSlo1* transcripts using qPCR measurements (Glauser, 2011).

The long reads generated by the Oxford Nanopore technologies (ONT) and Pacific Biosciences (PacBio) are used to assemble genomes either *de novo* or using hybrid approach [Goodwin et al., 2015; Koren et al., 2012]. These technologies are also suitable for sequencing full-length isoforms and provides information about exon connectivity and alternative splice site choices made at distal regions within each isoform. The application of NGS technologies to study exon connectivity is difficult in genes that can generate multiple different isoforms and, in these cases, the short-read data can only be used to infer isoform expression through statistical methods [Garber et al., 2011]. The CAMSeq [Sun et al., 2013] method developed to study alternative splicing in *Drosophila Dscam1* between exon clusters 4 and 9 involves multiple steps to reduce the template size from ~1.8 kb to ~1 kb and generates four reads per sequenced template. The SeqZip method developed by Roy *et al.,* [2015] uses DNA oligonucleotide-based

ligation that reduces the size of template by five times and does not require cDNA synthesis. While these approaches can utilize NGS methods to directly measure exon connectivity, they are prone to RT dependent template switching, ligation-based artefacts that can result in chimeric reads and require design of multiple ligamers specific to each target gene (SeqZip) and multiple circularization steps (CAMSeq).

Reverse Transcriptases (RT) catalyze the synthesis of complementary DNA (cDNA) from RNA and they are essential in the preparation of RNA-seq libraries. Retroviruses have been the primary source of RT enzymes, but these RTs are ineffective at synthesizing long, full-length cDNAs and stalling of these enzymes on RNA template can lead to prematurely terminated cDNA [Klarmann et al., 1992; Huber et al., 1989; Kotewicz et al., 1988]. RT enzymes commit both substitution and indel errors and approximately 1 in 17,000 and 30,000 nucleotides, respectively, were error prone in cDNAs synthesize by AMV and MMLV RTs respectively [Arezi & Hogrefe, 2006]. Many engineered RT mutants have been derived from retroviral sources. For example, SS3, can function at temperatures above 50^0 C but the error rates are not sufficiently low [Potter et al., 2003]. To improve the fidelity of cDNA synthesis, proofreading enzymes have been added to cDNA reactions that further reduced error rates by eight-fold [Arezi & Hogrefe, 2006]. For example, the addition of T4 bacteriophage gene 32 protein in the cDNA reaction improved the yields and synthesis of longer cDNAs [Villalva et al., 2001]. Recently, interest in RT enzymes coded by mobile group II introns obtained from bacterial sources have increased due to their higher processivity and fidelity compared to retroviral RTs [Zhao et al., 2018; Mohr et al., 2013].

The introduction of long-read sequencing technologies makes it possible to connect long distance exon information but the quality of full-length reads obtained from these technologies, depends partly, on the efficiency of RT used to generate the library. By combining the long-read platform with direct RNA sequencing technology, one can eliminate the need to rely on RT enzymes and quantitatively profile isoforms and measure their abundance without statistical

inferences. We have previously shown that the long-reads generated by ONT can be used to identify exon connectivity [Bolisetty et al., 2015] and in this current study, we utilize the ONT platform to assess its ability to sequence full-length cDNAs and whether it can be used in the expression profiling studies using Spike-In RNA Variant (SIRV) mixes. We additionally applied this technology to sequence ultracomplex genes from *Drosophila melanogaster* to identify near full-length isoforms. Further, we utilized the recently developed direct RNA sequencing technology to sequence yeast *Eno2* and compared the efficiency of full-length sequencing between direct RNA and cDNA sequencing experiments.

Results

Long read sequencing of E1 SIRVs on the MinION device

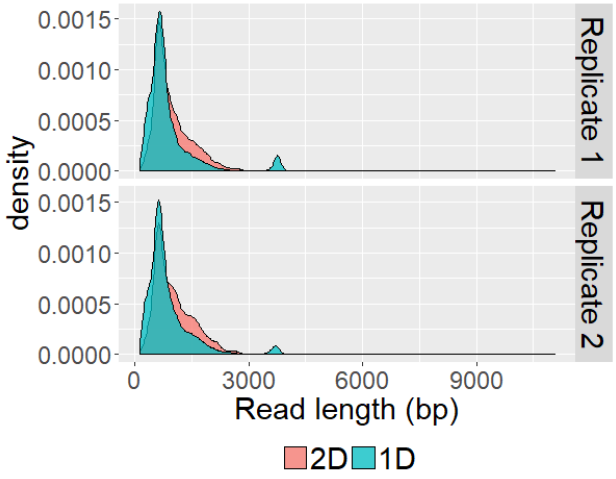
The E1 SIRV cDNA libraries generated using the MAP006 protocol were sequenced on the R7.3 flow cell and the replicates 1 and 2 were run on separate flow cells that each had over 500 and 700 pores available for sequencing respectively. Replicate 1 was sequenced for a total of 6 hours yielding 27.7 Mb and replicate 2 was run for 4 hours yielding 28 Mb. A total of 29,083 reads for replicate 1 and 29,391 reads for replicate 2 (Table 3.1) were obtained from the sequencing run and the read length histogram is shown in Figure 3.1A. The mean length for both replicates including both 1D and 2D reads were 955 bp and the read lengths for 2D reads were longer compared to 1D reads. We calculated the mean quality scores for each read and the scores corresponding to 1D and 2D reads are shown in Figure 3.1B. The mean quality score for 2D reads were higher than the 1D reads with a mean of ~10 and ~8 respectively but each replicate showed an average quality around 9.5. Next, we were interested in identifying whether there is any difference in sequencing quality with respect to read position. To address this question, the nanopores reads from the E1 SIRV libraries were divided into 10 equal bins and the mean quality

Table 3.1 Number of reads obtained from nanopore sequencing and LAST alignment statistics resulting from E1 SIRV spike-in experiment. The alignment results shown are for LAST tuned settings and the percent of aligned reads are shown within brackets.

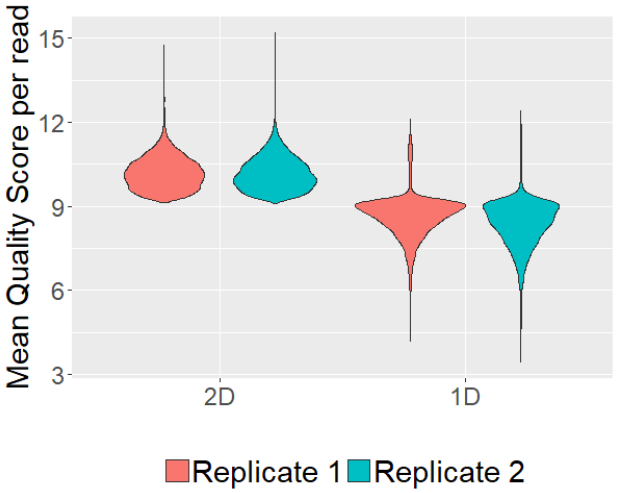
Sample	**Total Reads**	**2D reads**	**2D reads**	**2D reads**	**1D reads**	**1D reads**	**1D reads**
		No. of reads	Genome alignment	Transcriptome alignment	No. of reads	Genome alignment	Transcriptome alignment
E1 Replicate1	29083	20778	19924 (95.8%)	20301 (97.7%)	8305	6066 (73%)	6934 (83.5%)
E1 Replicate2	29391	14878	14360 (96.5%)	14473 (97.2%)	14513	11637 (80%)	12567 (86.6%)

Figure 3.1 A) Density plot showing the distribution of nanopore read lengths for 1D and 2D reads. B) Violin plot showing the mean quality of each read for both 1D and 2D reads C) Boxplot showing the mean quality of reads separated into 10 equal bins. Data for both 1D and 2D reads are combined.

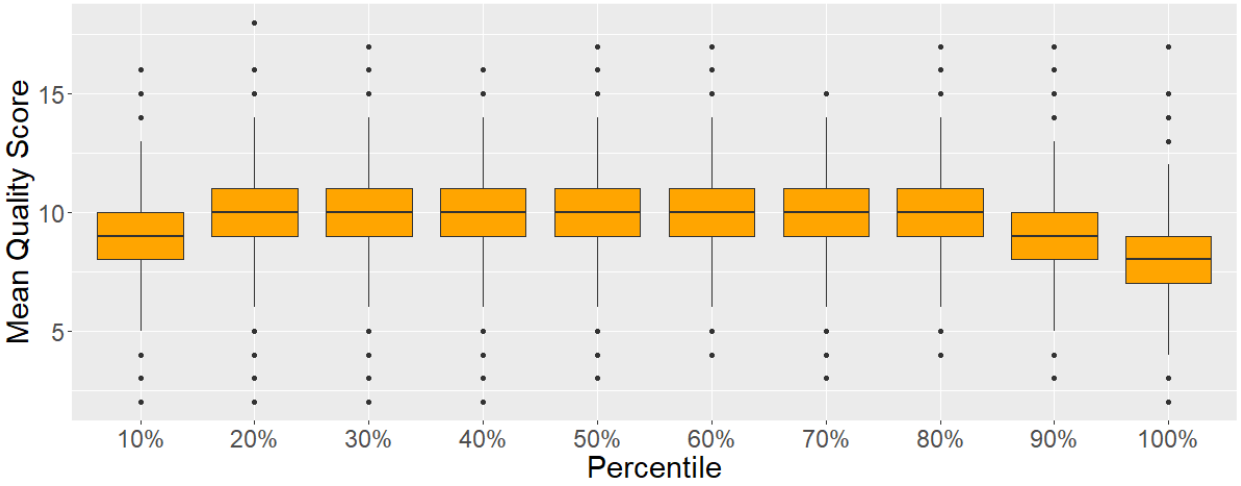
A


0.0015
0.0010
0.0005
0.0000
density
Replicate 1
Replicate 2
0
3000
6000
9000
Read length (bp)
2D
1D

B


Mean Quality Score per read
15
12
9
6
3
2D
1D
Replicate 1
Replicate 2

C


Mean Quality Score
15
10
5
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Percentile

score for each bin was calculated (Figure 3.1C). The boxplot revealed that the quality of bases did not vary much and the mean quality across all bins was found to be 9.7. We observed a slightly lower quality in the start and end regions and especially and the bin 10 had a mean quality of 8.3.

Aligning nanopore reads to the SIRV transcriptome using LAST

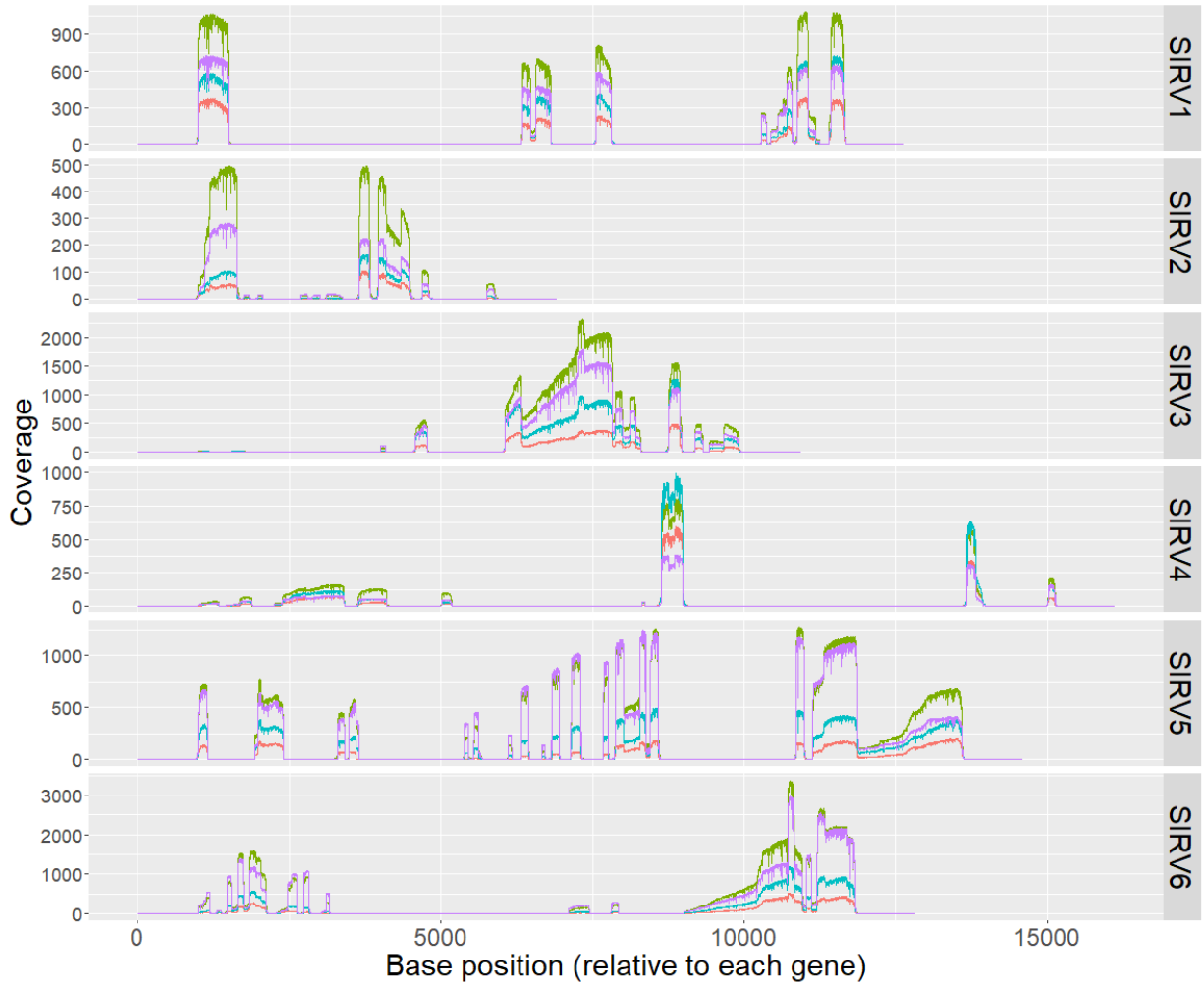
We aligned the E1 SIRV nanopore reads to the SIRV genome sequence and plotted the coverage for each SIRV gene (Figure 3.2). We observed a larger reduction in coverage at each base position in SIRV2, SIRV3 and SIRV6 and when we examined the orientation of these three genes, we found that 4 out of 6 SIRV2 transcripts were in the negative strand, and 16 out of 18 SIRV6, and 7 out of 11 SIRV3 transcripts were in the positive strand. The coverage is biased towards the 3' region and while the 5' regions show reduced coverage. This pattern is consistent with both read types in each replicate suggesting that the lower coverage is systematic. Next, we aligned replicate 1 reads to the SIRV genome using BWA-MEM [Li, 2013], GraphMap [Sovic et al., 2016], LAST [Frith et al., 2010] and Minimap2 [Li, 2017] programs to evaluate the aligner performance. We used the default options for all the above aligners and additionally used tuned parameters for LAST [Bolisetty et al., 2015]. The alignment results were analyzed based on the total number of reads aligned by each aligner and the unique versus multiple alignments for each read. Our analysis showed that LAST with tuned parameters aligned over 90% of the reads compared to other aligners (Figure 3.2B). GraphMap aligned a slightly lower number of reads compared to the default LAST but all the reads were found to be uniquely aligned with GraphMap. We chose LAST with tuned options for further analysis on the basis of its ability to align large number of reads and it has been previously reported to be the most inclusive aligner tool [Jain et al., 2015].

Table 3.2 Number of uniquely and multi-aligned reads with and without last-split option for E1 SIRV spike-ins.

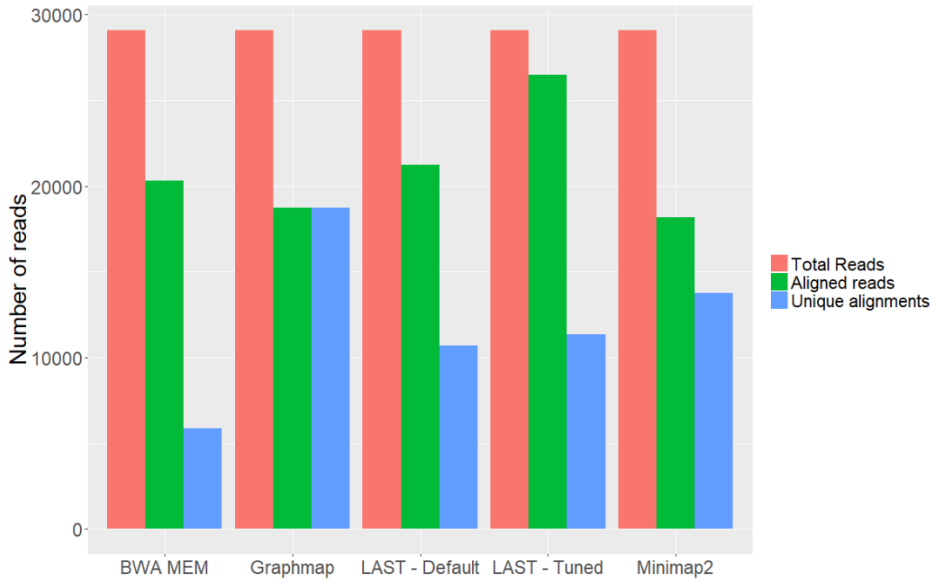
Sample	**Genome (with last-split)**	**Genome (with last-split)**	**Transcriptome (with last-split)**	**Transcriptome (with last-split)**	**Transcriptome (without last-split)**	**Transcriptome (without last-split)**
	Unique	multi-aligned	Unique	multi-aligned	Unique	multi-aligned
E1 Rep1 (2D)	8716	11208	19160 (94.3%)	1141	780	19521 (96.1%)
E1 Rep1 (1D)	3825	2241	6650 (95.9)	284	515	6419 (92.5%)
E1 Rep2 (2D)	5666	8694	13972 (96.5%)	501	398	14075 (97.2%)
E1 Rep2 (1D)	6412	5225	12252 (97.5%)	315	733	11834 (94.1%)

Figure 3.2 Genome coverage plot for SIRV spike-ins obtained from nanopore sequencing. A) The line graph showing the coverage for SIRV1 through SIRV6 and is color coded by replicates and read quality. B) Bar plot showing the total number of reads aligned and uniquely aligned reads for E1 SIRV replicate 1 using different aligners.

A


SIRV1
SIRV2
SIRV3
SIRV4
SIRV5
SIRV6
Coverage
Base position (relative to each gene)
0
5000
10000
15000

B


Number of reads
30000
20000
10000
0
BWA MEM
Graphmap
LAST - Default
LAST - Tuned
Minimap2
Total Reads
Aligned reads
Unique alignments

Next, we aligned these reads to the SIRV transcriptome and compared the results to the genome alignments. We found that over 95% of the 2D reads were aligned to reference genome and transcriptome whereas the 1D reads showed a slightly reduced percentage of reads aligned (Table 3.1). The alignment to the transcriptome resulted in slightly larger number of aligned reads compared to the genome alignments.

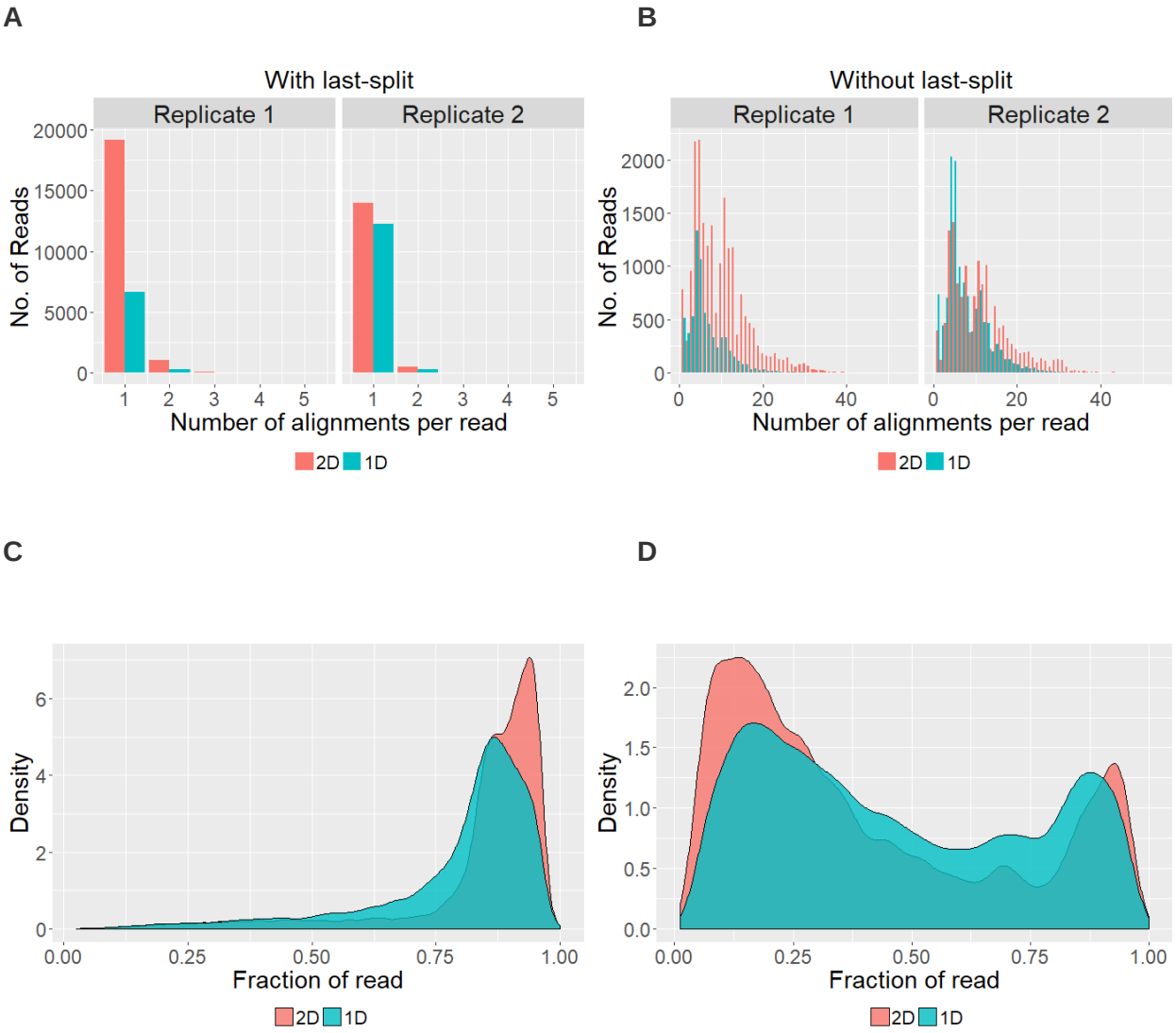
From the transcriptome alignment, we calculated the number of reads aligned to each isoform and since each SIRV gene contains multiple isoforms, we expected that the nanopore reads could be aligned to multiple isoforms. We first aligned the reads to the transcriptome without using the last-split option and compared it to the alignment with last-split option. We found that a large number of reads had multiple alignments without last-split with a median of 8 split-alignments per read (median: 2D reads - 9; 1D reads - 6) (Figure 3.3B). When alignments were done with last-split, the number of split-alignments reduced sharply and over 94% of reads had unique alignments (Table 3.2, Figure 3.3A). This is in contrast to the alignments without last-split where over 92% of reads had split-alignments (Table 3.3) and the large number of split-alignments were due to the ability of the LAST algorithm to find all possible alignments. Since the SIRV isoforms are highly identical to each other, the sequences corresponding to constitutive exon regions present in nanopore reads can be aligned to all reference isoforms containing that exon.

We next asked whether the reads obtained from SIRV experiments can be aligned from end to end and to address this question, we calculated the fractional alignment defined as the ratio of length of the aligned region to the total read length. When the last-split option was used, over 75% of reads from both replicates had a fractional alignment over 0.8 but when last-split was not used, only 18% of reads had fractional alignment over 0.8. We observed a bimodal peak at both lower and higher fractional alignment without last-split (Figure 3.3D) and the mean read for reads with fractional alignment of less than 0.25 is 1.2 kb. The longer read length with lower

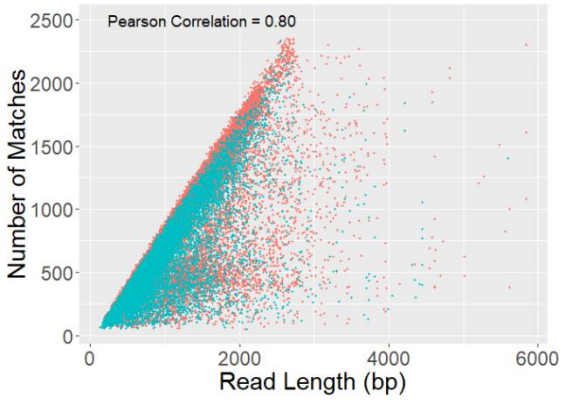
Table 3.3 Number of reads assigned to SIRV isoforms and the total number of isoforms manually assigned under different alignment thresholds.

Sample	**Total reads aligned**	**Total transcripts aligned**	**Total reads assigned**			**Total isoforms assigned**		
			25%	**50%**	**75%**	**25%**	**50%**	**75%**
Replicate 1	20301	67	5504	4852	3495	50	49	46
Replicate 2	14473	67	3721	3323	2560	49	49	47

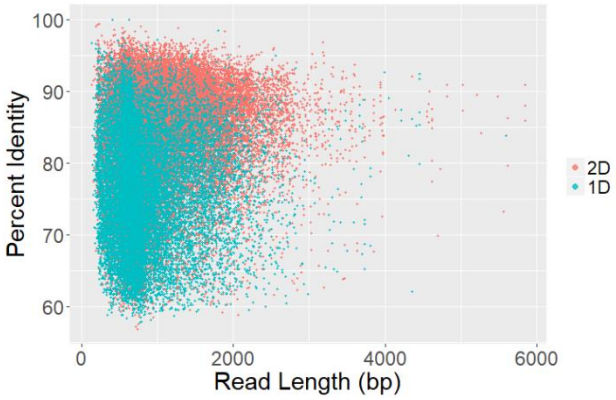
Figure 3.3 Comparison of read alignment using LAST with and without the last-split option. A, B) Bar plot showing the number of unique alignments for each nanopore read with (A) and without (B) the last-split option. C, D) Density plot showing the fractional alignment for each nanopore read measured as the number of bases aligned to the total length of the read. E) Scatterplot showing the correlation between the read length and the number of matches for each nanopore read reported by the LAST aligner. F) Scatterplot showing the percent identity for each aligned portion of the read. Panels E and F represent combined data from both replicates. G) Boxplot showing the number of matches per read reported by LAST for each SIRV isoform. The x-axis shows SIRV isoforms in the order of increasing size from left to right.


A
With last-split
Replicate 1
Replicate 2
No. of Reads
Number of alignments per read
2D
1D
B
Without last-split
Replicate 1
Replicate 2
No. of Reads
Number of alignments per read
2D
1D
C
Density
Fraction of read
2D
1D
D
Density
Fraction of read
2D
1D

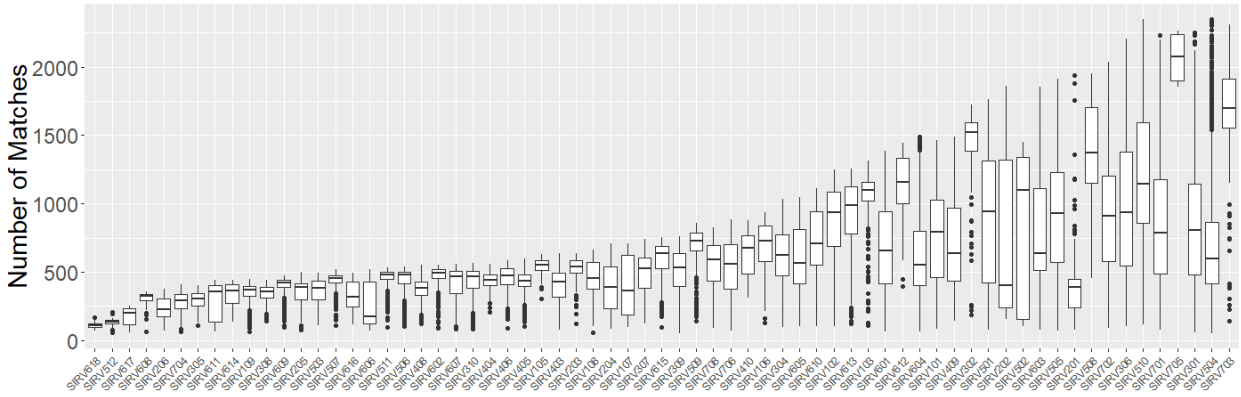
E


Pearson Correlation = 0.80
Number of Matches
Read Length (bp)
2D
1D

F


Percent Identity
Read Length (bp)
2D
1D

G


Number of Matches

fractional value indicate that these reads are not being aligned end to end to the reference. We found a strong correlation (Pearson R = 0.8) between the number of matches per read compared to its read length (Figure 3.3E). We next plotted the number of matches per read for all reads aligned to each SIRV isoform (Figure 3.3G) and found that as the length of the reference increased, the variation in the number of matches per aligned read increased.

Assigning of full-length nanopore reads to individual SIRV isoforms

To quantify the difference in abundances between various SIRV genes and isoforms, it is critical to have full-length reads that can be aligned to the full-length of the reference. Both the genome coverage (Figure 3.2 A) and fractional read alignment (Figure 3.3C) plots show that not all alignments reported by LAST are full-length. We randomly chose SIRV101 alignments for further examination and identified the bias in coverage in the 5' region (Figure 3.4B). We were therefore interested in identifying reads that span the full-length of the reference transcript and manually parsed the alignment information to identify full-length from non-full-length reads and assigned each read to its isoform of origin and the alignments with last-split option were used in further analysis. Since the last-split option resulted in fewer multi-aligned reads (Table 3.3), we first identified the best match for each multi-aligned read using the highest number of matches and retained only unique alignments for each read. Following this, the coordinates from the transcriptome were converted to genome coordinates. Using the transcript annotation file, each SIRV gene was separated into multiple segments based on the overlapping exon and intron regions and a unique identifier is generated for each isoform (Figure 3.4A). Next, for each isoform within a given gene, the LAST reported alignments were parsed to identify whether the alignments span each smaller region of the gene we identified in the previous step and thus, an identifier is created for each nanopore read. In the final step, the unique identifier for the reference is

compared against the aligned reads and reads were assigned to the reference if both identifiers match perfectly.

In generating the identifier for nanopore reads, we used different alignment thresholds (25%, 50% and 75%) to determine whether the alignments span a given gene region at least to this extent. The distribution of exon lengths from the SIRV transcript annotation showed that the median size is 125 bp with SIRV503 having the shortest exon length of 9 bp and SIRV504 having the longest exon length of 2.47 kb. The above approach to separate an isoform into multiple segments generated shorter gene segments and the median length changed to 84 bp and, in this case, by requiring a 25% threshold, a long read with at least 21 bp in this region will be called as having full-length alignment, and a 50% threshold requires 42 bp and a 75% threshold requires 63 bp. We chose three thresholds as a way to increase the stringency and for regions with less than 5 bases, we required a 100% alignment threshold to avoid mis-assigning reads. Using the above three thresholds, we were able to assign full-length reads to between 46 and 50 isoforms compared to the 67 isoforms reported by the LAST aligner (Table 3.3). A representative example of our approach to assign full-length reads and separate the non-full-length reads using SIRV101 is shown in Figure 3.4B. We found that as we increased the threshold, fewer reads were assigned and this is due to the stringent threshold that would filter out reads that are either misaligned or only partially aligned (Figure 3.4C). From these assigned reads, we calculated the fractional alignment for each full-length read to determine whether we observe any difference and the density plot in Figure 3.4D shows that the number of reads with a fractional alignment over 0.7 increased post-assignment compared to pre-assignments (Figure 3.4D). We further examined the distances between alignment start (dTSS) and end (dTES) coordinates corresponding to the first and last exons for the reference isoform and the alignment. We calculated dTSS and dTES for both assigned and unassigned reads and the boxplot in Figure 3.4E reveals larger differences between assigned and unassigned reads and the distances were strikingly larger for dTSS. A

similar pattern is observed for isoforms from all SIRV genes across all alignment thresholds (data not shown).

Next, we asked whether the replicates correlated well with one another and to address this question, we calculated transcripts per million (TPM) for each isoform. The scatter plot between replicates for 2D reads assigned at the 50% threshold showed good concordance with a Pearson correlation of 0.965 (Figure 3.4G). We further assessed whether the reads obtained from the MinION were able to quantitate the input RNA. The SIRV transcripts present in the E1 mix contains four sub-mixes each containing between 12 and 21 transcripts (Lexogen GmbH). Each sub-mix contains an equimolar concentration of RNA but the sub-mixes vary by 8 orders of magnitude. We identified the corresponding SIRV isoforms and grouped them into four sub-mixes, and the observed TPM and expected relative concentration difference between sub-mixes were plotted (Figure 3.4H). The long reads generated by nanopore contains sequencing errors and as shown in figure 3.3B for SIRV101, the insertions (purple) and deletions (black) can be easily observed. As the errors are distributed randomly, reads that overlap a given region can be used to correct these errors and the corrected reads have been used on nanopore-only generated data to assemble bacterial genome *de novo* [Loman et al., 2015]. We were interested in identifying whether a similar approach can be used to correct nanopore reads from our SIRV experiments. To address this, we identified the SIRV isoforms that have at least 50 reads assigned and used the correct module in the Canu [Koren et al., 2017] assembler. We were able to correct random errors and the median percent identity improved from 85.7% to 96.9%, pre- and post-correction, respectively [Figure 3.3H].

Figure 3.4 Assigning full-length nanopore reads to SIRV isoforms. A) The UCSC genome browser track customized to show the SIRV genome exemplifies our approach to assign full-length reads to isoforms from individual genes. The top half of the track shows the GTF annotation for SIRV1 and the bottom half shows how distinct bed regions are created for each SIRV1 isoform using overlapping regions. The black colored regions in the track represent exons and the grey regions represent introns. For SIRV101, the track is expanded to show how each exon from annotation is further divided into smaller regions that are used in the assignment process.

B) IGV browser track showing all LAST reported alignments to SIRV101. The top two tracks show the coverage for replicates 1 and 2 respectively. The third track shows BAM alignments for replicate 1 and the purple color shows insertions and the black color shows deletions. The coverage plot below the bam track shows the assigned and unassigned reads respectively using replicate 1.

C) Bar graph showing the number of full-length nanopore reads assigned to their corresponding SIRV isoforms at different alignment thresholds.

D) Density plot showing the fraction of nanopore reads aligned to the reference for both assigned and unassigned reads.

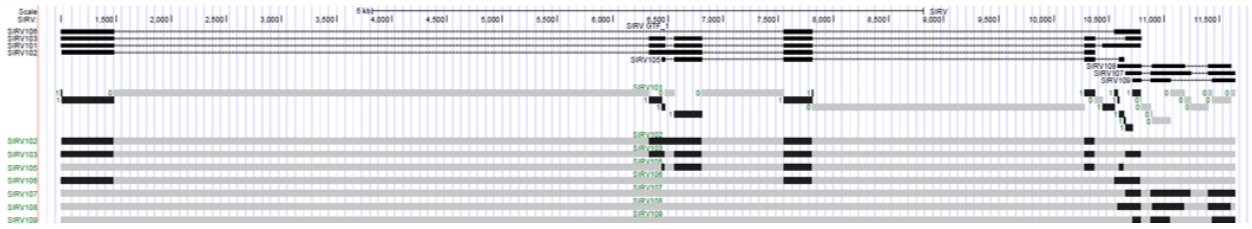
E) Boxplot showing the distances between reference and alignment start (dTSS) and reference end and alignment end positions (dTESS).

F) Violin plot showing the improvement in percent identity between pre- and post-correction using Canu correction module.

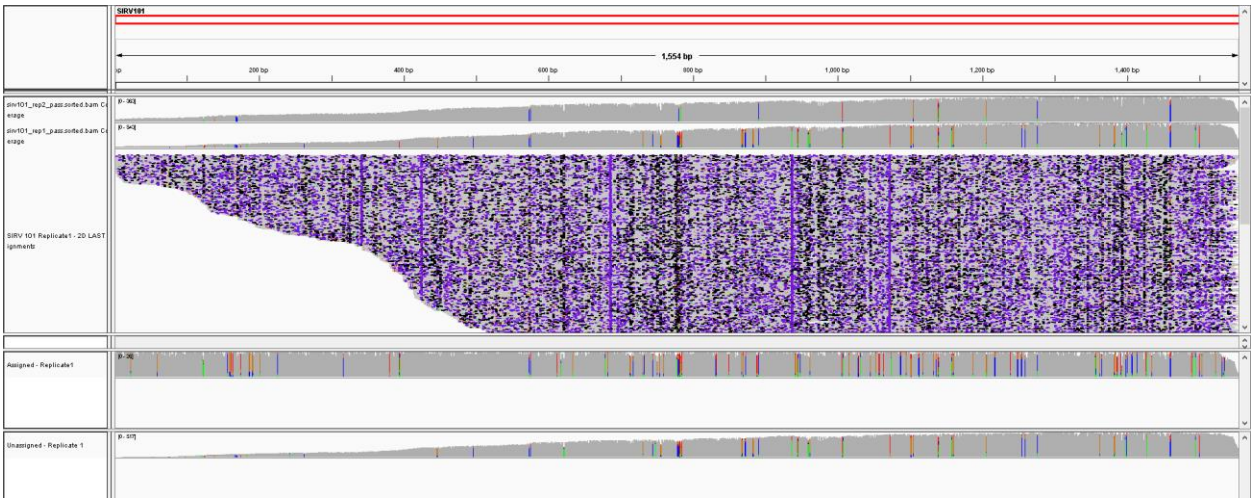
G) Scatterplot showing the correlation between E1 SIRV replicates based on the number of assigned reads using 50% alignment threshold. The Pearson correlation is 0.965.

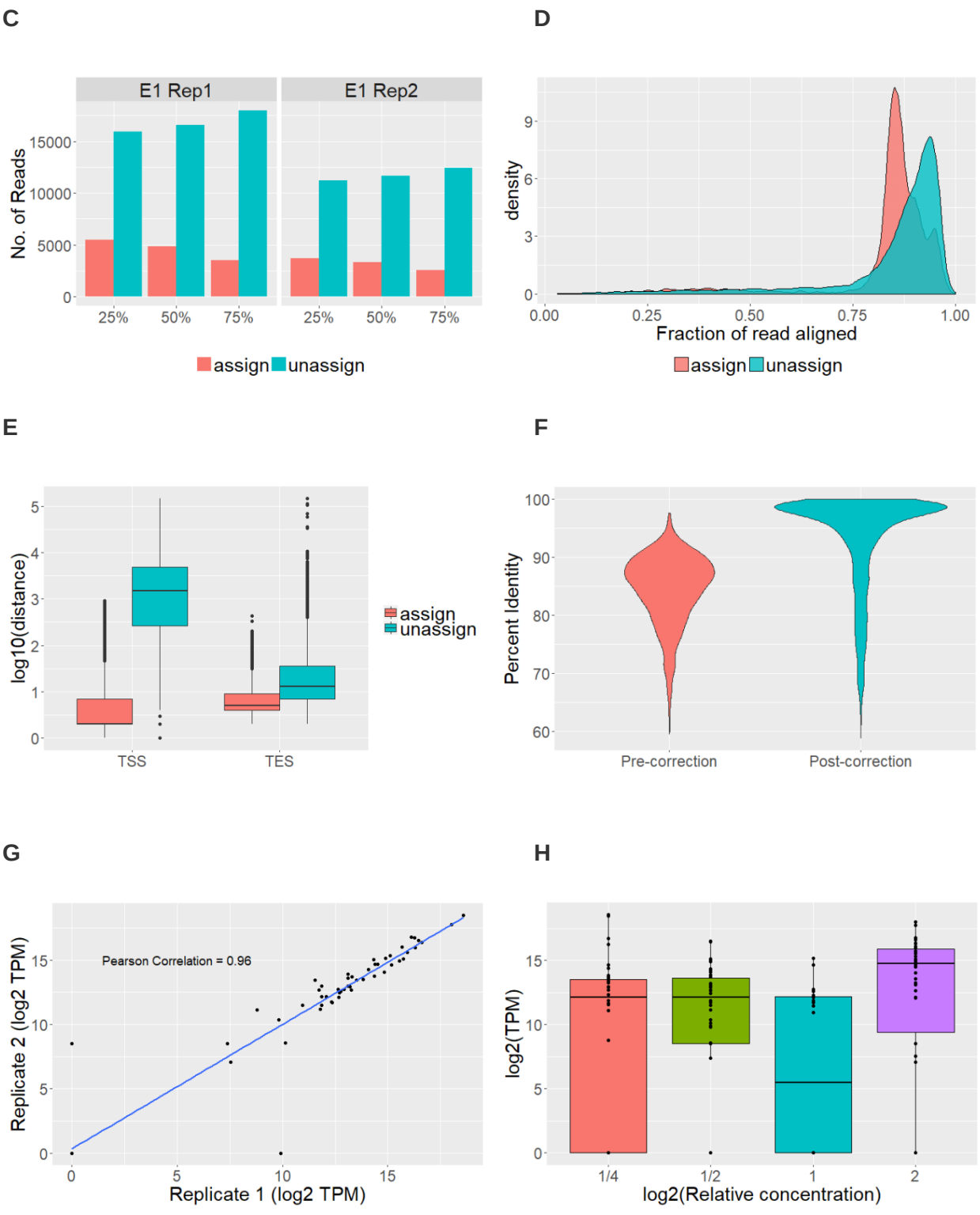
H) Boxplot showing the log2 transformed TPM based on the assigned reads compared to the relative concentration of input RNA present in four sub-mixes. The x-axis represents log2 fold change in the sub-mix concentration and the sub-mixes 1 through 4 varies by 8 orders of magnitude. Data for both replicates are combined.

A



B


SIRV101
1,554 bp
Assigned - Replicate1
Unassigned - Replicate 1


C
E1 Rep1
E1 Rep2
No. of Reads
15000
10000
5000
0
25%
50%
75%
assign
unassign
D
density
9
6
3
0
0.00
0.25
0.50
0.75
1.00
Fraction of read aligned
assign
unassign
E
log10(distance)
5
4
3
2
1
0
TSS
TES
assign
unassign
F
Percent Identity
100
90
80
70
60
Pre-correction
Post-correction
G
Pearson Correlation = 0.96
Replicate 2 (log2 TPM)
15
10
5
0
Replicate 1 (log2 TPM)
H
log2(TPM)
15
10
5
0
1/4
1/2
1
2
log2(Relative concentration)

Long-read amplicon sequencing of ultracomplex genes in *Drosophila*

After establishing a method to assign full-length reads generated by the ONT platform, we were interested in sequencing *Drosophila* genes that were identified to undergo complex splicing patterns based on multiple splicing events across the gene body. The RNA-seq data from adult head samples [Graveley et al., 2011] were analyzed further and we identified 11 genes (*Aldh-III, app, alph, Fur1, gish, Ndae1, PMCA, rdgB, Sap47, Sgg,* and *SK*) that undergo seven different types of alternative splicing events and were termed as ultracomplex genes (UCGs). The modENCODE MDv3 annotation [Brown et al., 2014] contained a total of 27,314 isoforms (Table 3.4) for these 11 genes, 4 of which can generate over 1000 isoforms each. To reduce the total number of isoforms in our sequencing experiments, we limited our primer design to transcripts that were annotated in the Flybase and RefSeq annotations in the UCSC genome browser. The primer binding sites (Table 3.5) were present in a total of 19,984 isoforms but only a total of 5,916 unique isoforms (Table 3.4) can be identified from these UCGs and their transcript sizes range from 917 bp to 14.8 kb.

Following library generation, we loaded the amplicons on the MinION flow cell - 536 pores were available during platform QC but only 454 were available during sequencing. The samples were run for around 2 hours and after basecalling, we used Poretools to extract fasta reads and obtained a total of 20,844 2D and 9,728 1D reads. The median read length for 2D reads were 1.9 kb whereas the 1D reads were 0.7 kb. We aligned the 2D reads to the reference transcripts using LAST and obtained 15,350 reads (73.6% alignment rate) of which 12,609 reads (60.5%) had unique alignments with a mean fractional alignment of 0.91. LAST aligned these reads to a total of 679 isoforms from all 11 UCGs and in the case of *Fur1*, all seven isoforms were reported to contain alignments with a total of 902 reads. The mean percent identity of these unique alignments was 79.7% and we used the above approach to assign full-length reads to UCG isoforms. By using a 50% alignment threshold, we assigned 3,269 reads representing approximately 26% of

the uniquely aligned reads to 161 isoforms (Figure 3.5A; Table 3.4). We were able to assign reads to only one isoform for *rdgB* and *Sap47* but *gish*, *Aldh-III*, *PMCA* and *SK* had over 10 isoforms each (Table 3.4). We further explored the full-length reads assigned to *Fur1* because of the lower number of isoforms and our ability to assign 6 out of 7 isoforms. We calculated the edit distance between *Fur1* isoforms and found that *Fur1-RE* and *Fur1-RC* were only different by 21 bases and we were able to distinguish between these two isoforms using this 21 bp region on the UCSC genome browser (data not shown).

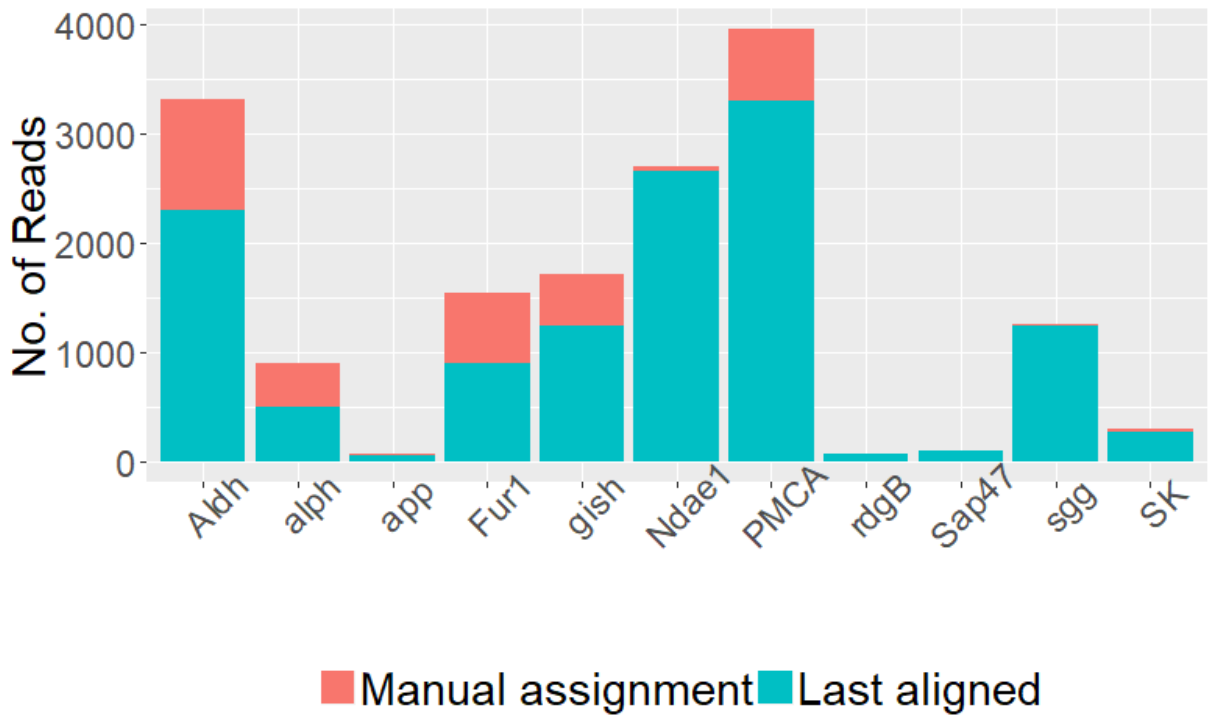
Additionally, we were interested in examining full-length assigned to *PMCA* because the PCR amplification resulted in visible products from only one of the two primer sets. For *PMCA*, we assigned 659 reads to 15 isoforms and five isoforms were assigned with over 50 reads each (Figure 3.5B). During library preparation we observed PCR bands from one set of primers (Fwd2 + Rev1; see Methods & Materials) but we pooled all PCR reactions together to generate libraries. We identified 14 isoforms that map to primer set 1 (Fwd1 + Rev1) and interestingly, we found one nanopore read that was assigned to *PMCA-uw* that can only be generated by primer set 2 (Fwd 2 + Rev1; Figure 3.6A). The reads that were not assigned to *PMCA-uw* were examined further and found that the unassigned reads did not pass the 50% alignment threshold in the 5'UTR region (Figure 3.6B). We analyzed other *PMCA* isoforms and found that two isoforms contained exons not previously annotated in either RefSeq or Flybase (Figure 3.6C) but were present in the MDv3 annotation (Brown et al., 2014). In addition, we also found a region in the *SK* gene where a homopolymer region showed a reduced coverage compared to nearby regions (Figure 3.6D) and the 10 full-length *SK* isoforms are shown in Figure 3.6E.

Table 3.4 Total number of isoforms for 11 ultracomplex *Drosophila* genes obtained from the MDv3 annotation [Brown et al., 2014]. The number of unique isoforms that can be identified from the primer designs and the reads aligned by LAST and manual assignments are shown.

Genes	Number of isoforms present in the MDv3 annotation	Number of isoforms within by primer sequences	Total number of isoforms aligned by LAST	Total number of isoforms assigned
gish	18972	3122	265	79
Aldh-III	359	195	80	33
PMCA	632	63	51	15
SK	1117	620	98	10
Fur1	7	7	7	6
app	22	22	14	5
alph	30	11	8	5
Ndae1	389	194	76	4
sgg	369	100	25	2
rdgB	406	256	26	1
Sap47	5011	1326	29	1
Total isoforms	**27,314**	**5,916**	**679**	**161**

Figure 3.5 Assigning full-length reads to *Drosophila* ultracomplex genes. A) Bar plot showing the total number of reads aligned by LAST-tuned settings to different UCG isoforms (blue) and the number of reads manually assigned at 50% alignment threshold (red). B) Bar plot showing the number of full-length reads assigned to isoforms from 8 different UCG. The x-axis shows the number of isoforms and is orders from highest to lowest based on the number of reads manually assigned.

A


No. of Reads
4000
3000
2000
1000
0
Aldh
alph
app
Fur1
gish
Ndae1
PMCA
rdgB
Sap47
sgg
SK
Manual assignment
Last aligned

B

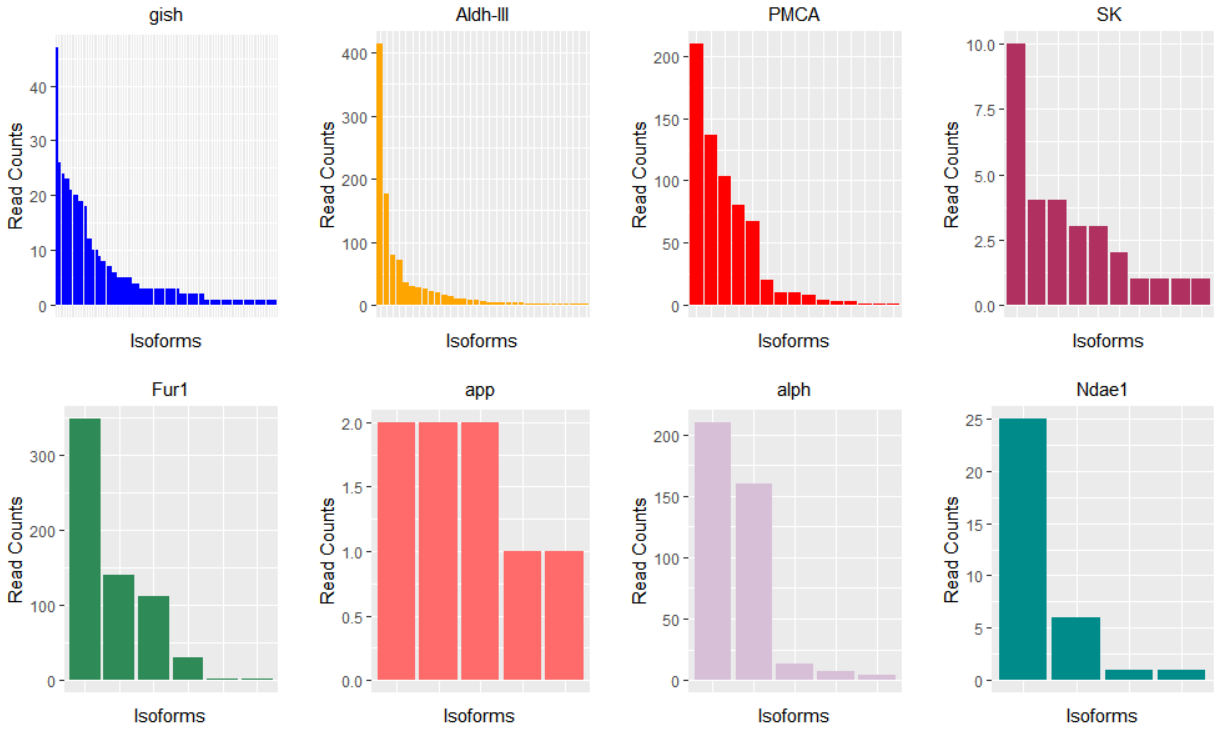

gish
Aldh-III
PMCA
SK
Fur1
app
alph
Ndae1
Read Counts
Isoforms

Figure 3.6 UCSC genome browser track showing full-length reads assigned for PMCA isoforms.

A) UCSC genome browser track showing *PMCA-uw* isoform that were amplified by primer set 2 The primer binding sites are shaded in blue, yellow and orange and the MDv3 annotation corresponding to this isoform is shown in black color. The red color track shows the read assigned to *PMCA-uw* and the blue colored PSL track shows the unassigned reads.

B) The unassigned reads were enlarged to show that they lack the 50% alignment threshold. The 5' UTR region (length = 52 bases) is shown and the shaded yellow region represents the 26 bases.

C) The nanopore read that was assigned to two other isoforms (red and grey colored tracks) containing exons not annotated in RefSeq (yellow) and Flybase (blue) is shown.

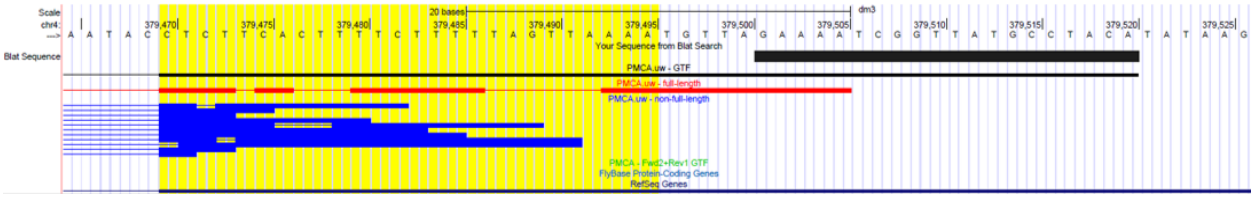
D) IGV browser track showing a homopolymer region in the *SK* gene shown in the red colored bars for base T. A large number of deletions is observed in this region shown by black horizontal lines compared to adjacent regions and the coverage plot on top reflects this with reduced coverage.

E) UCSC genome browser track showing all 10 *SK* isoforms that were assigned with full-length nanopore reads. The yellow and orange shaded vertical bars represent the forward and reverse primer sites and the PSL tracks below show different *SK* isoforms (color-coded) that are full-length. The expanded track for *SK.egin* isoform (tan color) shows the unassigned reads that either lacks sufficient alignment in the 5' or 3' exon regions or only partially covers the isoform. The PSL track in red, immediately above, shows the full-length reads assigned to *SK.egin*.

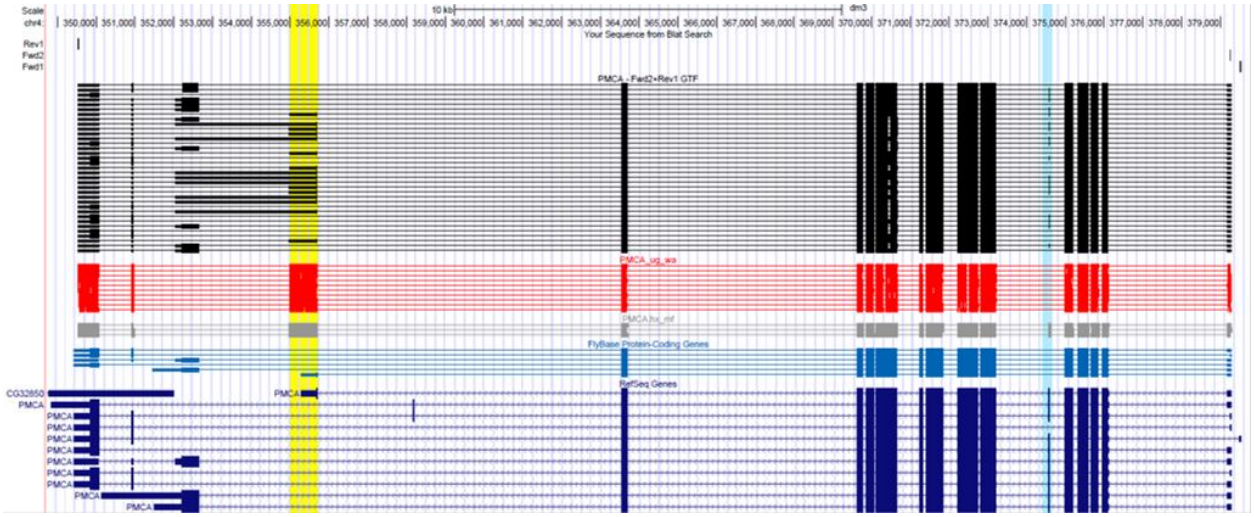
A


Scale
chr4:
10 kb
dm3
Your Sequence from Blat Search
Rev1
Fwd2
Fwd1
PMCA.uw
PMCA.uw - GTF
PMCA.uw - full-length
PMCA.uw - non-full-length
PMCA - Fwd2+Rev1 GTF
FlyBase Protein-Coding Genes
RefSeq Genes
CG32850
PMCA

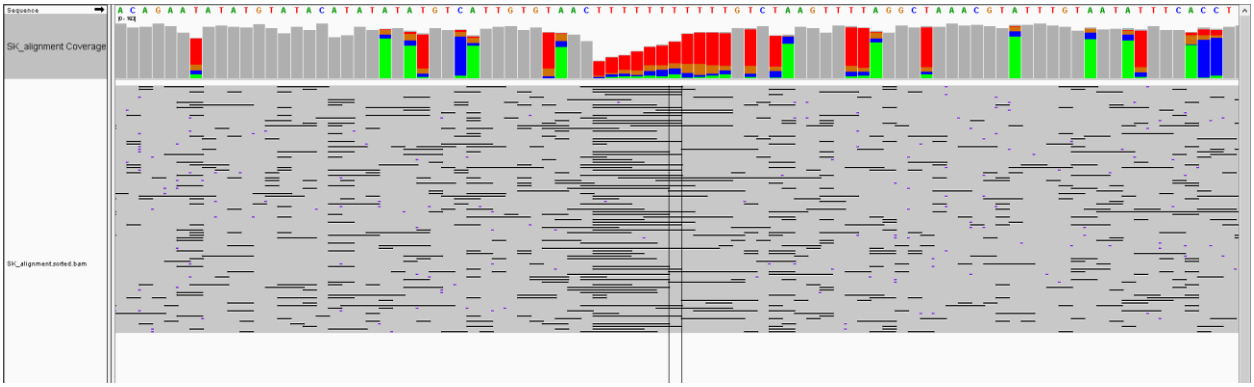
B


Scale
chr4:
20 bases
dm3
Blat Sequence
Your Sequence from Blat Search
PMCA.uw - GTF
PMCA.uw - full-length
PMCA.uw - non-full-length
PMCA - Fwd2+Rev1 GTF
FlyBase Protein-Coding Genes
RefSeq Genes

C


Scale
chr4:
10 kb
dm3
Your Sequence from Blat Search
Rev1
Fwd2
Fwd1
PMCA - Fwd2+Rev1 GTF
FlyBase Protein-Coding Genes
RefSeq Genes
CG32850
PMCA

D



E

RNA editing in the ultracomplex genes

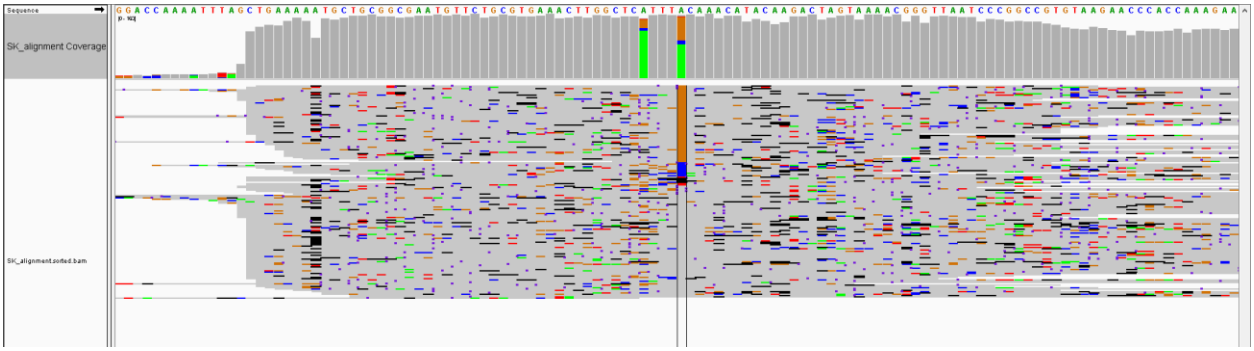
Next, we were interested in examining the RNA editing sites present in the ultracomplex genes we sequenced. We searched for transcript locations previously reported by Graveley *et al.,* (2011) to undergo RNA editing and identified a total of 6 sites present in *SK* (1 site), *Sap47* (1 site), *sgg* (2 sites) and *Ndae1* (2 sites). Of these 6 sites, 4 has the potential to change amino acids but two of these sites present in *Sap47* and *sgg* result in silent changes. We first examined all the reads aligned to the *SK* genomic loci at chrX:5290552 (dm3 build) and we were able to identify the RNA editing (Figure 3.7A). The IGV track shows the A>G mismatches supported by 159 reads, 54% of which contained the reference A base and 37% contained G whereas C and T were supported by 7% and 2% of the reads, respectively. Similarly, we were able to identify the A>G SNPs supporting from *Ndae1* (Figure 3.7B, C). However, the lower coverage (<100 reads) at the RNA editing site present in *sgg* and *Sap47* did not allow us to distinguish between A>G mismatches and sequencing errors. When we compared the RNA editing sites present in *SK* and *Ndae1* to the adjacent regions, we found these errors to be systematic and this suggests that a larger proportion of mismatches we observed at these sites are unlikely to be sequencing errors and could potentially represent *bona fide* editing sites. Since *Ndae1* contains 2 editing sites that are separated by >1 kb in the genome, we further evaluated the combinatorial RNA editing for each *Ndae1* isoform. Since the maximum number of reads that were assigned for the top *Ndae1* isoform is 25 reads, we were not able to distinguish between sequencing errors and RNA editing but the alignment tracks sorted by base at site 2 in *Ndae1* suggests combinatorial editing at these two sites.

Figure 3.7 IGV tracks showing Adenosine-to-Inosine RNA editing sites from *Drosophila* ultracomplex genes.

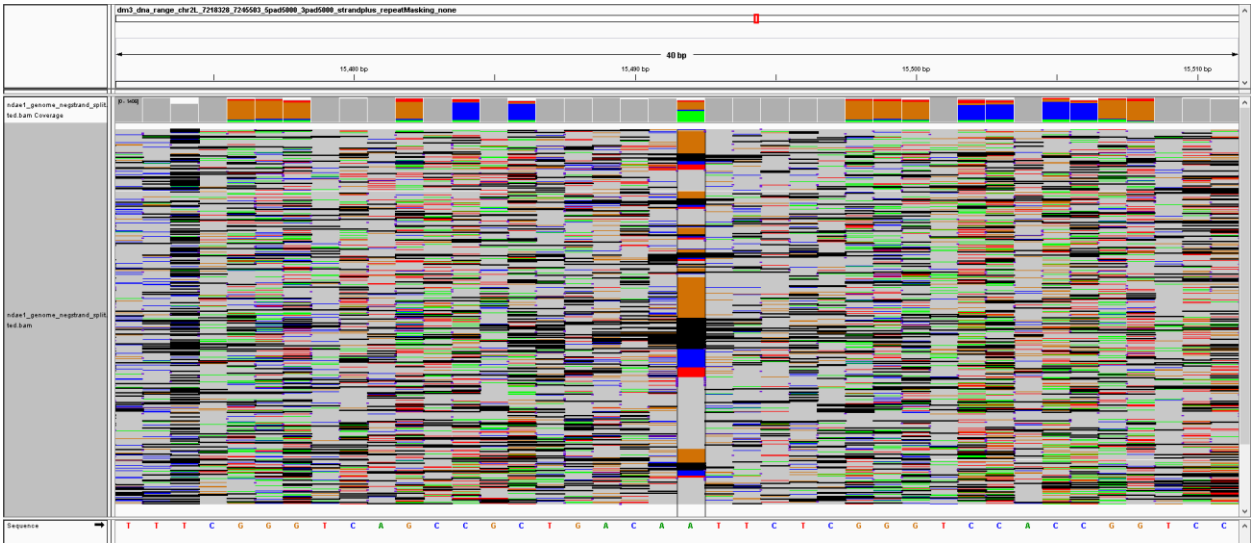
A) The RNA editing site identified in *SK* gene using nanopore sequencing that results in A to G mismatch is shown in the center of the IGV track. In the alignment track, A is shown in Green, T in Red, C in Blue and G in Brown. The BAM alignments are sorted by base and the mismatches at each position are shown in colors corresponding to each base and the black color represents deletion. The allelic frequency is set at 0.2 and the coverage plot on top shows the color-coded proportion of mismatched bases at any position above this threshold.

B and C) IGV tracks showing the RNA editing sites present in *Ndae1* transcripts at site 1 (B) and site 2 (C). The edited sites are shown in the center and the A to G mismatches are shown in Brown color. The bam alignments are sorted by bases at site 2 and the corresponding editing is shown in site 1. The coverage at each of these sites vary but panels B and C show combinatorial RNA editing.

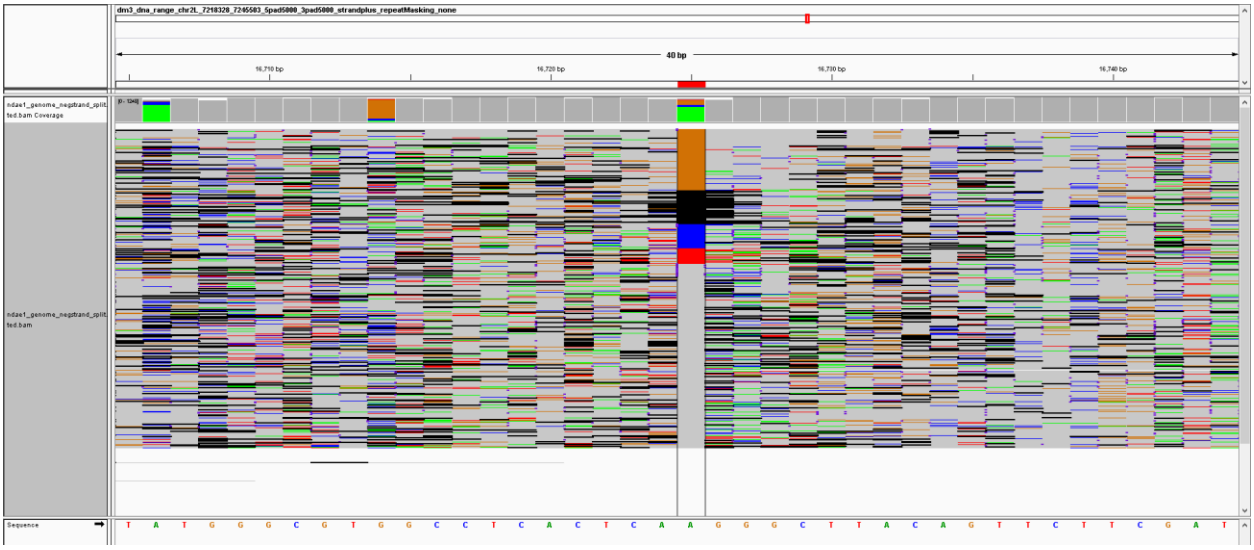
A


Sequence
SK_alignment Coverage
SK_alignment.sorted.bam

B


dm3_dna_range_chr2L_7218328_7245583_5pad5000_3pad5000_strandplus_repeatMasking_none
40 bp
15,480 bp
15,490 bp
15,500 bp
15,510 bp
ndae1_genome_negstrand_splitted.bam Coverage
ndae1_genome_negstrand_splitted.bam
Sequence

C


dm3_dna_range_chr2L_7218328_7245503_5pad5000_3pad5000_strandplus_repeatMasking_none
40 bp
16,710 bp
16,720 bp
16,730 bp
16,740 bp
hdae1_genome_negstrand_split
ted.bam Coverage
hdae1_genome_negstrand_split
ted.bam
Sequence
T A T G G G C G T G G C C T C A C T C A A G G G C T T A C A G T T C T T C G A T

Direct RNA sequencing of yeast *Enolase2*

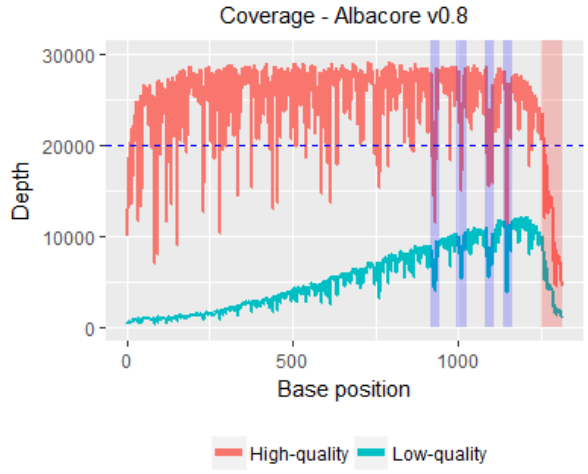
Currently, the ONT platform offers the ability to conduct direct RNA sequencing experiments without the need for synthesizing cDNA. While RT enzymes are widely used in NGS experiments, the retroviral RTs are not efficient at synthesizing full-length cDNA and the most widely used SuperScript enzyme shows uneven coverage and 3' bias [Mohr et al., 2013; Zhao et al., 2018]. By directly sequencing the RNA molecules, cDNA synthesis using reverse transcriptases can be entirely eliminated. To test whether direct RNA sequencing is capable of sequencing full-length RNA molecules and to identify if there is any bias in coverage, we sequenced yeast *Enolase2* (*Eno2*) RNA on the MinION and obtained a total of 48,177 reads (median length of 1.1 kb), of which 27,333 were high quality and 20,844 were low quality reads. We aligned the high-quality reads to the *Eno2* reference and obtained alignments for 27,145 reads (99.3% of high quality reads) with a median fractional alignment 0.96 showing that a large number of bases within each read can be aligned to full-length of the reference.

Next, we looked at the nanopore read coverage across *Eno2* to see if there is any bias. The median coverage across the entire transcript was 26,383 for high-quality reads but the low-quality reads shows bias in the 3' region and decreased dramatically from the 3' to 5' end with Albacore v0.8 (Figure 3.8A). We also observed few regions where the coverage dropped below 20,000 and was found to be present in both low and high-quality reads as shown by the shaded region (Figure 3.8A). The 3' end of *Eno2*, in particular, had very low coverage with Albacore v0.8 but basecalling with Albacore v2.2 showed improvements in coverage (Figure 3.8C). We were further interested in identifying the base composition between low and high coverage regions between these basecallers and we used 20,000 reads as the threshold, ~25% below the median with Albacore v0.8. Examination of base composition revealed that low coverage regions in both basecaller versions had a higher percentage of T's (Figure 3.8B). While Albacore v2.2 showed bias against C's and T's, there were only 54 regions that fall below 20,000 reads whereas with

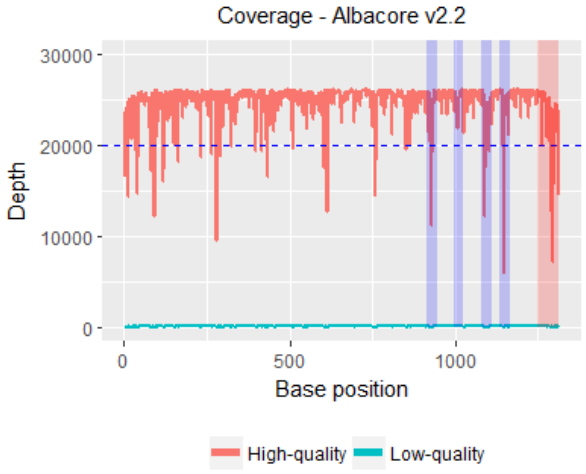
v0.8, 158 regions had lower coverage. These results show that Albacore v2.2 improves the basecalling with only few regions showing bias. To compare differences between direct RNA sequencing and cDNA sequencing, we identified five SIRV isoforms (102, 103, 601, 610, 613) that range in size from 1.15 kb to 1.42 kb and were size matched with *Eno2*. We calculated the fractional alignment from these two experiments and found that cDNA sequencing had a mean of 0.87 compared to 0.95 for direct RNA sequencing experiments.

Figure 3.8 Direct RNA sequencing experiments using yeast *Eno2*. The basecalling was performed with Albacore v0.8 (A) and Albacore v2.2 (C). The high-quality reads are shown in red and the low-quality reads in blue. The shaded regions in blue indicate regions with lower coverage in both low and high-quality reads. The dashed line represents the cut-off used to perform base-composition analysis B) Bar plot showing the difference in base composition in the *Eno2* with Albacore v0.8 and D) with Albacore v2.2. The high coverage represents positions where the coverage is above 20,000 and below this threshold is defined as low coverage. The baseline represents coverage across the entire region.

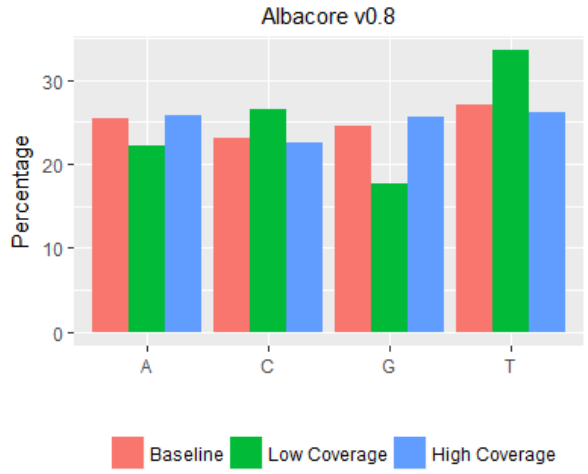
A


Coverage - Albacore v0.8
30000
20000
10000
0
Depth
0
500
1000
Base position
High-quality
Low-quality

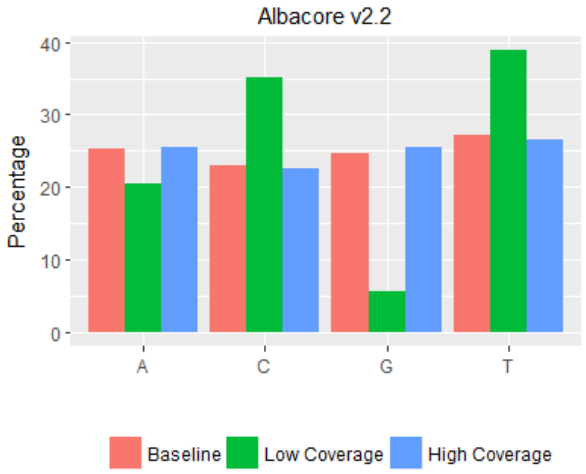
B


Coverage - Albacore v2.2
30000
20000
10000
0
Depth
0
500
1000
Base position
High-quality
Low-quality

C


Albacore v0.8
30
20
10
0
Percentage
A
C
G
T
Baseline
Low Coverage
High Coverage

D


Albacore v2.2
40
30
20
10
0
Percentage
A
C
G
T
Baseline
Low Coverage
High Coverage

Discussion

In this study, we utilized the ONT platform to sequence full-length cDNA libraries generated from SIRV spike-ins. The reads obtained were separated into low and high quality and the mean quality score for the SIRV dataset was approximately 10 indicating that the overall accuracy for the nanopore reads were 90% and is reported to be similar from other published works [Oikonomopoulos et al., 2016; Ip et al., 2015]. The higher error rates inherent to nanopore reads causes a large proportion of reads to be unaligned [Oikonomopoulos et al., 2016] and we observed a similar trend in our analysis with multiple aligners. LAST was found to be the most inclusive of all aligners [Jain et al., 2016] and we observed similar results with tuned options in our sequencing experiments (Table 3.2) but the 1D reads still returned fewer alignments compared to 2D reads.

The isoforms from SIRV genes contain overlapping exons and the non-full-length reads can be aligned to multiple isoforms. The split-alignments reported by LAST poses challenge to study isoforms that are highly identical to each other [Bolisetty et al., 2015] but we were able to manually parse the alignment information and assign full-length reads to their isoforms of origin. We assigned between 17% and 27% of full-length reads in this study and while the proportion of assigned reads is very low, they represent high confidence assignments. The 75% threshold can be stringent and can result in classifying reads to be unassigned if the threshold is not met and we found this to be the case with *PMCA-uw* isoforms. The main advantage of the long-reads generated by nanopore sequencing is that no statistical inference is necessary to measure isoform abundances.

Comparison of total number of reads assigned with high-confidence to the expected input RNA concentration shows wide variation in sub-mixes 1 and 3. The disagreement between observed and expected concentration we observed in our experiments could be explained by the large number of non-full-length reads and can affect the number of reads assigned to each

isoform. The cDNA synthesis protocol we followed should ideally enrich for full-length cDNAs but we cannot exclude the possibility that incompletely synthesized RNAs were also selected in the linker ligation step. During cDNA synthesis, if reverse transcriptases pauses sites are near G, the cap-dependent linker ligation protocol can also enrich for these non-full-length cDNAs but not all non-full-length reads are found to end with G's. By extending our read assignment approach to ultracomplex genes from *Drosophila*, we were able to assign reads to 161 isoforms using amplicon sequencing. Since UCGs used in this study contains multiple alternative exons, single molecule long-read technologies can be used to study exon connectivity. The long reads obtained from PacBio and ONT have been shown to study genes that undergo complex splicing events [Bolisetty et al., 2015; Treutlein et al., 2014]. Comparison between spike-in cDNA and direct RNA sequencing experiments show that the size-matched amplicons from SIRVs show lower mean fractional alignment. But the direct RNA sequencing showed lower coverage in regions that were enriched for T's and this bias in homopolymer basecalling is reported in prior studies [Jain et al., 2018]. The low-quality reads show poor coverage in the 5' region of the reference but the high-quality reads do not show this trend.

In summary, the data we present here provides evidence that the ONT platform can generate both full-length DNA and RNA reads. Our experiments were conducted during the early stages of ONT development and we obtained low throughput as a result of loss of actively sequencing pores. The constant developments in technology and increased throughput with PromethION and MinION with latest R9.4 chemistry can be used to quantitate isoform expression and deconvolute isoform connectivity. Further improvements in basecallers using Scrappie and Albacore can provide even coverage in homopolymer regions.

Materials and Methods

Spike-in RNA variant mix library preparation

The E1 SIRV library was prepared as provided in the instruction manual for Teloprime kit from Lexogen GmbH. We synthesized cDNA using 2 µl of E1 SIRV RNA (25.2 ng/ul) in a 20 µl reaction at 46 °C for 50 mins. The cDNA reaction was purified using the silica column provided in the kit and adapter ligation was performed at 25 °C for 3 hours and was subsequently purified and eluted in 14 ul RNA buffer. The second strand was synthesized by mixing 13 µl of cDNA with second strand and enzyme mix and cycled as follows: 98 °C for 90 s, 62 °C for 60 s, 72 °C for 5 mins. The double stranded cDNA is purified and eluted in 20 µl DNA buffer and 9 µl of E1 SIRV cDNA template was used in the end-point PCR as follows: 98 °C for 45 s; 50 °C for 90 s; 72 °C for 5 mins; 98 °C for 30 s, 62 °C for 60 s, 72 °C for 5 mins (39 X); 72 °C for 5 mins; 10 °C hold. The fluorescence plateaued at 22 cycles and we used 19 cycles at ~80% fluorescence in the downstream library generation steps. To prepare replicate 1 library, 9 µl of ds cDNA was PCR amplified for 19 cycles and for replicate 2, a separate cDNA reaction was set and 9 ul of ds cDNA was amplified for 19 cycles. The PCR products were Ampure purified and we retained 345 ng and 400 ng for replicate 1 and 2 respectively.

Amplicon library preparation for Drosophila ultracomplex genes

For the library preparation involving ultracomplex genes, multiple individual cDNA and PCR amplifications were required. The fly heads from Bloomington stock 2057 were used to extract total RNA using Trizol (Sigma) reagent and between 1 and 2 ug of total RNA was used in each cDNA reaction with SS II (Invitrogen) using oligo$(dT)_{20}$VN primer. The first strand was synthesized at 47 °C for 50 mins and 2 µl cDNA was used in a 30 µl LongAMP PCR protocol using gene specific primers as shown in table 3.5 below. The primers shown below were designed

based on the Flybase and RefSeq annotation and the genes containing only one reverse primer (as in the case of *Aldh-III*) were combined with the forward primers and vice versa. For genes containing multiple forward or reverse primers, we first checked a given primer set can amplify multiple isoforms and if did, the extension time for PCR was based on the longest isoform such that we can amplify all possible isoforms amplified by a given primer set. Each PCR reaction was done for 25 cycles and we used melting temperature provided by Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/ as a starting point and if a reaction did not amplify the cDNA, we tested multiple conditions. In addition, for *gish* and *PMCA*, we were not able to design primers due to the presence of repeat sequences.

Library preparation for direct RNA sequencing using *Eno2* RNA

Th library for direct RNA sequencing was prepared using SQK-RNA-001 kit as per the ONT manufacturer instructions. The RNA CS containing poly-A+ *Enolase2* was used as the starting material and 9.5 ul was ligated with the RT adapter in a 25 µl reaction for 15 mins at room temperature. After ligating adapter, Superscript IV (Invitrogen) was used to synthesize first strand cDNA at 50 °C for 10 mins followed by inactivation of enzyme at 80 °C for 10 mins. The RNA:cDNA hybrids were purified using 1.8X Ampure beads in DNA LoBind tubes and eluted in 20 ul ultrapure water. RNA adapter was ligated using T4 DNA ligase at room temperature for 10 mins and purified using 1X RNA AMpure beads and eluted in 21 ul manufacturer provided elution buffer. The eluate contained 180 ng final library as measured by Qubit.

Nanopore sequencing of amplicons from SIRV and ultracomplex genes

For SIRV library preparation, the MAP006 protocol was used and 345 ng of replicate 1 and 400 ng of replicate 2 was end repaired and A tailed using NEBNext UII kit and incubated for

20 °C for 5 mins and 65 °C for 5 mins. The end-repaired/A-tailed DNA was Ampure purified (1.8X) and the HP adapter was ligated using Blunt/TA ligase (NEB) at room temperature for 10 mins. To this reaction mixture, 1 µl of HP tether was added and incubated at room temperature for 10 mins. MyOne^{TM} Streptavidin C1 (Invitrogen) beads were used to purify the adapter ligated/tethered library and this pre-sequencing library was eluted in 25 µl ONT supplied elution buffer.

The sequencing library was prepared by mixing 3 µl pre-seq mix with fuel mix and RNB buffer and 150 µl final library mix was loaded on R7.3 flow cell and each replicate were sequenced in separate flow cells. The library was replenished by loading more sequencing mix as necessary.

For ultracomplex genes, all PCR amplified reactions were pooled together and Ampure purified and eluted in 100 ul ultrapure water at a concentration of 18.9 ng/µl. Using the MAP006 protocol, 1 µg of purified amplicons were end repaired and A-tailed using NEBNext UII kit as described above. Following Ampure purification, amplicons were ligated with adapter and tethered, and purified with MyOne C1 beads. The pre-seq mix was eluted in 25 µl elution buffer and 6 µl pre-seq was mixed with fuel mix and RNB buffer before sequenced on R7.3 flow cell. The raw reads obtained from SIRV and ultracomplex genes were basecalled using Metrichor and Poretools v0.3.0 [Loman & Quinlan, 2014] was used to extract fasta reads.

Direct RNA sequencing of *Eno2*

For direct RNA sequencing, 100ng of the final library made from *Eno2* was mixed with RBF1 and loaded on the R9 flow cells and the raw reads were basecalled and sequences were extracted using Albacore. We base called the raw reads and fasta sequences were extracted using Albacore v0.8.4 and v2.2.0 [Oxford Nanopore Technologies].

Table 3.5 List of primers used in the *Drosophila* ultracomplex sequencing experiments

Gene name	Type	Sequence	Tm
Aldh-III	Fwd1	CTCGCGCTTATTAGCCAACT	59.65
	Fwd3	GTTGTCAAAGCGCGTCAACT	61.41
	Rev	TCACATCACCCACCGATATG	60.2
alph	Fwd1	AGGCGATGAAAATCAAGTGG	60.07
	Fwd2	GTCATCCGCTGGTGAAAGTT	60.12
	Rev1	TTCAAGTTGTGTGGCTCTGC	60.03
	Rev2	GCCATCCATTCCATCATTTC	60.1
Fur1	Fwd	GAGTTCTGCGGTGGAAAATC	59.68
	Rev1	GTATCTGTGGTGGTGCGTGT	59.47
	Rev2	GCCCAAAGATAGCTGCTGAC	59.98
	Rev3	CCTTCTCATCAGCTCGCCTA	60.64
	Rev4	GAGCTGTAATGCCGTCGAAT	60.24
gish	Fwd1	AAATCAGCCGCAAAAACAGT	59.75
	Fwd2	AATCGCACCATTTTTCTCGT	59.57
	Fwd3	CTCCGCCGTATCATTGTTTT	59.96
	Rev1	TCTCTCTTGGATGTGGGTAAGG	60.49
	Rev2	AATGTGTACGGGGTTCATTTG	59.6
app	Fwd1	TCGGGTTCACTCCAATTTTC	59.91
	Fwd2	TGCTCTTTCGATTGTGTTTCC	60.24
	Fwd3	ATAACGCAACGGCACTCATT	60.53
	Fwd4	GTGTTGCTGTCATGGCTTTG	60.31
	Rev1	GCAATGGAGCTTGTCGAAAT	60.22
	Rev2	GATCGAGTTGAGATCGATGTTG	59.7
	Rev3	AAGCAAGCAGTGTGCTCAGA	59.93
Ndae1	Fwd1	GTCAAACCATAACCAAGCCAAT	60.12
	Fwd2	GGGCACAGAGCTCTCAACTT	59.6
	Fwd3	ATCGCGTTTTTCAACCGAAT	61.7
	Rev1	CGCATGATATGTTGCCAGAC	60.1
	Rev2	GCCTCCTTGCAATTGTTGAT	60.08
	Rev3	TGGTTGTGTTCAATGGTTGG	60.25

PMCA	Fwd1	TGTAGGCATAACCGATTTTC	55.34
	Fwd2	GAAAAGGAAGGCGTCACAGT	59.33
	Rev1	ACCGATCCGCTTACATTTTG	59.96
rdgB	Fwd1	CAGCGCACGCTGTTTTT	59.71
	Fwd2	AGAACAACCGCTGACTGACC	60.31
	Rev1	CCATCAAACAACAGGAGGAAA	59.96
	Rev2	CAATTGGTTTGCTTGAGTTGC	60.66
	Rev3	TGGCATTCTCAATCAGTTGG	59.65
Sap47	Fwd1	GCGCAGTTGTTGTTTCCATA	59.74
	Rev1	GTTGTTAGTCGAGCGTGTGC	59.52
sgg	Fwd1	TCACGCTTTACAGTCGGACTAA	59.94
	Fwd2	AAAAGCAGCAGAGCGTGTTT	60.2
	Fwd3	CCGAAAACTTGGAAACCACT	59.07
	Fwd4	CCTTTGATACCCGAGTTTGC	59.57
	Fwd5	GGCCATCACCTTTTAGCCTTA	60.45
	Fwd6	CGCGACTCTATTTGCCTGTT	60.4
	Fwd7	TGTTGCACGCTGAGAAAAAC	60.03
	Rev1	TCTTTGCCCCCAAATAGCAT	61.67
	Rev2	TGCTGCTGCTGTTCTTTCAT	59.75
	Rev3	CCTTCGGGGATTTCTCTTTC	60.01
SK	Fwd1	TCCGTCCCGTTATATGCTGT	60.35
	Fwd2	TTGTACCAGAAATGCCATGC	59.55
	Fwd3	TTACACCCATGGCTGACGTA	60
	Rev1	TGATGGGCACAATGATGAGT	59.93
	Rev2	AGACGGACACGAGAATTTGC	60.2
	Rev3	ATCTCCCCTCGACTTCCATT	59.9

Aligning nanopore reads and data analysis

The fasta reads obtained from each experiment were separately aligned using LAST [Frith et al., 2010]. The transcript sequences using bedtools [Quinlan & Hall, 2010] from the genome sequence provided by Lexogen GmbH for SIRV experiments and from the dm3 build for ultracomplex gene experiments using the MDv3 annotation [Brown et al., 2014]. For the ultracomplex genes, the fasta sequences from transcripts were further trimmed to start and end with the primer sequences using custom scripts in R [R Core Team, 2018]. The reference index for each experiment were created using lastdb and the nanopore reads were aligned using −s 2 -T 0 -Q 0 -a 1 (tuned option) both with and without last-split option. The alignment files reported in the MAF format was further converted to SAM and PSL using maf-convert.py script in the LAST program. To assign full-length reads to their references, we used custom scripts made in R. Since we aligned the reads to transcriptome, we first converted the alignment coordinates from transcriptome to genome space. The annotation file containing all isoforms for a given gene was parsed and the overlapping exon regions were used segment a given gene into multiple smaller segments and these segments were used to create a binary identifier such that all isoforms in a given gene are unique (Figure 4A). In the next step, we checked whether each read reported to be aligned by LAST contains an alignment in a given region and this process was repeated for throughout the entire length of gene and was done for all aligned reads. After this process, a binary identifier for each read was created and if the unique identifier for the reference isoform is identical to the read, then a read is assigned and in all other cases, the reads were unassigned. For the ultracomplex genes, we additionally obtained gene regions from dm3 build and aligned all 2D reads using LAST.

For the direct RNA sequencing, the *Eno2* sequence was obtained from sacCer3 build from chrVIII:451327-452640 and the lastdb index was created for alignment with LAST using the tuned option and no last-split option was used. The resulting MAF alignments were converted to SAM

and PSL formats using maf-convert.py. Using samtools [Li et al., 2009], bam files were created and were visualized in IGV browser. To compare the differences between basecaller algorithms, the fasta reads from each version of Albacore were separately aligned to the reference and the genome coverage was plotted for both. The base composition at low and high coverage regions were obtained from *Eno2* from each position and the results were analyzed.

Comparison of different long-read aligners

To test the performance of different aligners, we combined the 1D and 2D reads from E1 SIRV experiments and aligned these reads using the default settings using Bowtie2, BWA-MEM, GraphMap, LAST and Minimap2. For LAST aligner, we additionally performed alignment with the tuned option. From the resulting SAM files, we measured sensitivity based on the total number of reads aligned and the total number of unique alignments.

Correcting nanopore reads using the Canu assembler

The full-length reads that were manually assigned to the SIRV isoforms were identified and separated from unassigned reads. The fasta reads corresponding to the assigned reads were separated from the unassigned reads and were used to correct the nanopore sequencing errors using the correction module in the Canu assembler [Koren et al., 2017]. A 100bp length was used in both minReadLength and minOverlapLength parameters. The resulting corrected reads were separately aligned to the corresponding reference isoforms using LAST-tuned option as described above except the -m was reduced from 1000 to 100. The percent identify was calculated from PSL files and all other figures in this research work were plotted using ggplot2 [Wickham, 2009].

CHAPTER 4

A Summary of Findings, Conclusions and Prospects for Future Research

In my doctoral dissertation project, I set out to find whether the long-reads generated by the ONT platform can be used to study alternative splicing. To address this question, I initially targeted a small region in four *Drosophila* genes *Rdl*, *MRP*, *Mhc*, and *Dscam1*, in the order of increasing complexity from 4 possible splice isoforms for *Rdl* to 19,0008 isoforms for *Dscam1*. I chose the mutually exclusive splicing regions in these genes to amplify and comprehensively analyze these regions. I examined the *Dscam1* isoforms between exon 3 and 10 clusters because the ~1.8 kb length between this region presented me with a challenge to directly identify how the alternative exons at cluster 4, 6 and 9 are connected and whether any alternative exon choice at any of these clusters are correlated with each other, and an understanding of this question has tremendous implications for *Dscam1* biology. All exon variants, except 6.11, in *Drosophila Dscam1* were found to be expressed using NGS technology and at different developmental time points. The short-reads obtained from NGS platforms do not tell whether two exon variants from adjacent clusters are connected together in the mature transcript. This problem is further compounded by the reverse transcription process by which cDNAs are synthesized. Since the *Dscam1* exon clusters 4, 6 and 9 are located at least ~5 kb away from the 3' end, an efficient cDNA synthesis is necessary to study full-length *Dscam1* transcripts but the RT enzymes lacking processivity makes it difficult to study longer transcripts. To address this problem and to study smaller regions encompassing the clusters 4 through 9, I synthesized cDNAs by designing primers that can bind at exon 10 for *Dscam1* and made it easier to synthesize cDNAs from the ~1.8 kb region. The *Dscam1* exon variants are highly identical to each other and the presence of incompletely synthesized cDNAs can generate chimeric DNA molecules during the PCR amplification.

I addressed this by including a control that enabled me to optimize PCR amplification needed to both obtain enough material for sequencing and simultaneously reduce the proportion of chimeric molecules. The *in vitro* transcribed *Dscam1* spike-ins also served as a control for other genes used in this study because the identity of exon variants in *Mhc*, *MRP* and *Rdl* are lower compared to *Dscam1*. I tested three different PCR amplification reactions from 20, 25 to 30 cycles and as expected, increasing the cycle number increased template switched products. In the 20 cycle reaction, I found 0.2% chimeric molecules but this number increased to over 30% with 30 cycle reaction. The 20 cycle PCR gave reduced yield of the libraries and I used the 25 cycle reaction which increased template switched products but produced higher library yield. From these experiments, I was able to identify 42% of the 18,612 *Dscam1* isoforms containing exon 4, 6 and 9 variants, and ~55% of these isoforms were represented by more than one read while the rest of the isoforms were represented by only one read. When I examined isoforms from other genes, I found only 12 out of 180 isoforms for *Mhc*, 9 out of 16 for *MRP* but all four isoforms were identified for *Rdl*. From the 301 reads obtained aligned to *Rdl*, I performed further analysis to identify whether the splicing of exon variants at two different clusters were dependent and I found evidence for independent splicing between clusters.

To extend this analysis further, I utilized the ONT platform to sequence SIRV spike-in mixes and asked whether this platform can be used to measure isoform abundances. To quantitate isoform abundances, it is essential to obtain full-length cDNAs at high coverage for all isoforms expressed by each gene. In my SIRV experiments, I obtained a total of ~60K reads from both replicates with a mean quality score close to 10. Since the 2D reads are of high quality and is generated from both template and complement reads, I analyzed 2D reads further and a comparison between different aligners showed LAST with its tuned settings aligned large number of reads but a large proportion of the reads were non-full-length and represented a problem for subsequent analysis even when the alignment was done with last-split option to generate unique

alignments for a majority of the reads. I separated full-length reads from non-full-length reads and assigned between 17.5% and 26.5% of the 2D reads. As I increased the stringency, the number of reads assigned were decreased and less than 50 out of 69 isoforms were identified from this data. The data from SIRV replicates correlated well with each other but few isoforms did not have any reads assigned and, in few cases, the reads could be assigned to one replicate and not the other. The transcripts present in E1 SIRV are grouped into four different concentrations which allowed me to compare the expected and observed abundances. I found that each of the four sub-mixes varied to a larger extent the lower coverage for each isoform did not allow a comprehensive evaluation of the quantitative ability of the platform.

In addition to the SIRV experiments, I conducted experiments on 11 *Drosophila* ultracomplex genes to characterize their isoforms. This experiment is similar to our previous study done with *Dscam1* and three other genes but the ultracomplex genes contain alternative exons that undergoes other types of splicing in addition to mutually exclusive splicing and the splice choices are made across the full-length of the transcript from 5' to the 3' end. The genome annotation file consisted of over 27,000 unique isoforms from these 11 genes and to be able to obtain a 10X coverage, I needed at least 270,000 full-length nanopore reads. During the time I conducted my experiments, the ONT technology was in the development phase and generating a 10X coverage was difficult to achieve and so I narrowed down my total number of transcripts to ~6000 based on the annotations from RefSeq and Flybase. The design of primers was also challenging because of the overlapping exons and a primer designed to bind this region will result in the synthesis of cDNA from all transcripts that share this exon region, provided they are all expressed. This was observed to be the case with one primer set designed for *Sap47* and I did not make libraries from these primers. Additionally, in some cases, few combinations of primers were enough (*Aldh-III*) were enough to capture all isoforms but in all other genes, multiple primers were necessary. Regardless of the different primer sets used, the approach I used here cannot

be used to study changes in alternative splicing beyond the primer binding sites. From this experiment, I was able to assign 15.6% of all 2D reads to 161 isoforms out of the 5,916 possible isoforms present in the MDv3 annotation. The proportion of assigned reads were low because many reads obtained in these experiments were non-full-length. Despite the higher sequencing errors and shorter read lengths obtained from this platform, I was able to identify RNA editing sites from multiple sites. By using the overlapping reads, I was able to improve the sequencing errors that are randomly distributed. The ONT technology has progressed very rapidly and currently direct RNA molecules can be sequenced in addition to sequencing DNA. Direct RNA sequencing eliminates the need to synthesize cDNAs from RNA and I used this sequencing technology to show that full-length reads from yeast *Eno2* can be generated and more importantly, a larger fraction of the sequenced reads can be aligned. While I found differences in coverage between different Albacore basecaller version, the version 2.2 showed less bias and is better at calling homopolymer sequences.

RNA pull-down to enrich isoforms

In the experiments with *Drosophila* ultracomplex genes, I used primers that were specific to different isoforms of a gene. One of the limitations of this approach was that the primer design was based on the UCSC genome browser annotation but the study conducted by Brown et al., (2014) identified novel isoforms from these genes. Isoforms with multiple start and end sites make it difficult to characterize isoforms in a comprehensive manner because the primers need optimization. As mentioned earlier, multiple isoforms can contain overlapping exons and unique primers cannot always be designed. To profile all isoforms of a gene, probes can be designed across the constitutive exons to capture all expressed isoform. The transcripts captured this way can be used in the cDNA synthesis that either attaches the 5’ adapters using strand switch method using SmartSeq2 protocol or the cDNAs can be separately ligated with 5’ adapters. The group II

intron reverse transcriptases (TGIRT or Marathon RT) can be used during the cDNA synthesis step because these enzymes are processive and can generate full-length cDNAs.

Identifying novel isoforms

The method I used herein to assign full-length reads use overlapping exon regions between multiple isoforms from the annotation file. A curated set of annotation is a prerequisite for this approach to characterize the transcriptome. I applied this method to study *in vitro* synthesized SIRV spike-ins and few *Drosophila* ultracomplex genes. These two experiments represent a relatively simple case because the composition and concentration of different transcripts present in the SIRV mix is precisely known and a complete annotation is possible. The biological samples represent a complex case where the isoforms of a given gene can have varying abundances and in addition, a complete annotation is often not possible. By using an incomplete annotation, novel isoform combinations cannot be identified using the above approach I used to study SIRV and ultracomplex genes. In these cases, the long-reads can be first aligned to the genome instead of transcriptome to identify splice sites and all possible transcripts can be generated. In the next step, the reads can be aligned to the newly generated transcript annotation and by assigning reads to each isoform, novel isoforms can be identified and quantified.

Sequencing ultra-long transcripts

The isoforms from *Drosophila* ultracomplex genes I studied range in size from 0.9 kb to ~15 kb. Since the read length obtained from nanopore is only limited by the read length present in library, ultralong transcripts that are over 20 kb in length can be studied. From the MDv3 annotation, I found over 115,000 isoforms from 37 genes are over 20 kb in length and are suitable candidates to be sequenced on the ONT platform. Efficient synthesis of full-length cDNA from

Figure 4.1 Histogram of isoform lengths obtained from *Drosophila* MDv3 annotation [Brown et al., 2014]. The x-axis represents the bins of isoform lengths and the y-axis represents the counts of isoforms within each bin. The numbers above each bar represents the total number of genes in each bin.

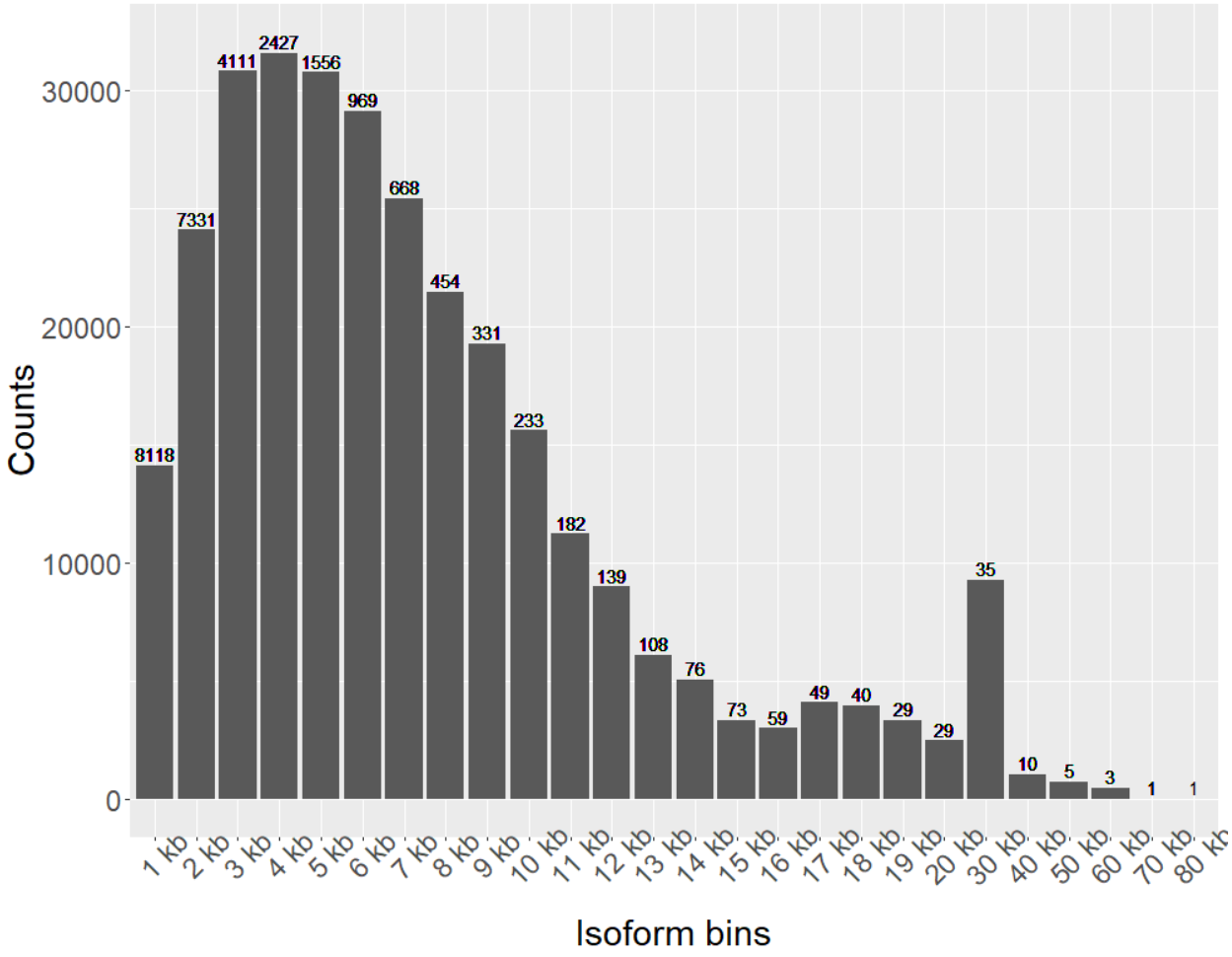
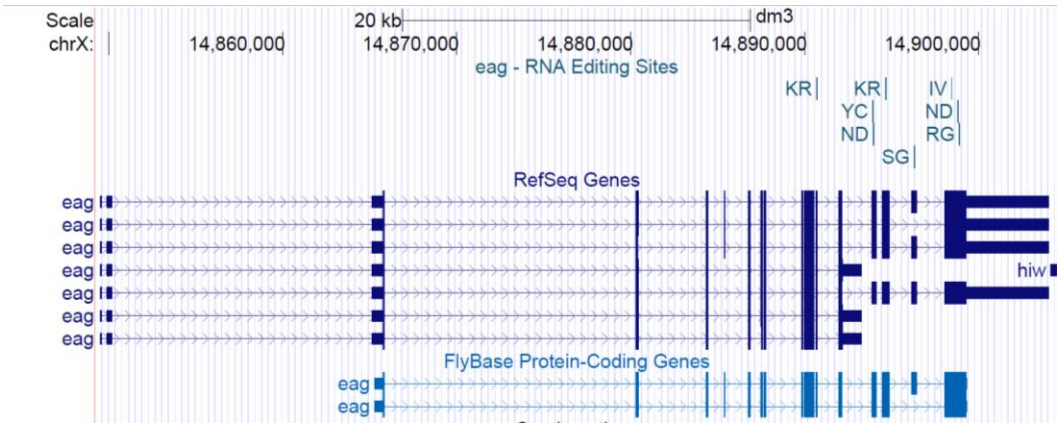
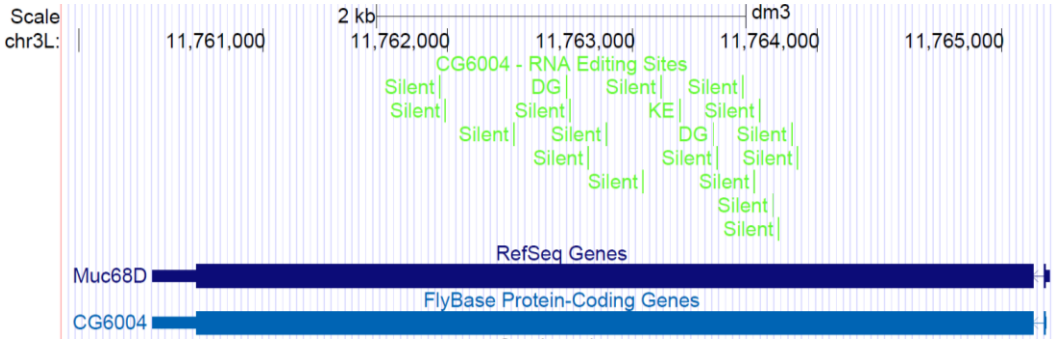

Counts
30000
20000
10000
0
8118
7331
4111
2427
1556
969
668
454
331
233
182
139
108
76
73
59
49
40
29
29
35
10
5
3
1
1
1 kb
2 kb
3 kb
4 kb
5 kb
6 kb
7 kb
8 kb
9 kb
10 kb
11 kb
12 kb
13 kb
14 kb
15 kb
16 kb
17 kb
18 kb
19 kb
20 kb
30 kb
40 kb
50 kb
60 kb
70 kb
80 kb
Isoform bins

Figure 4.2 Examples of *Drosophila* genes with multiple RNA editing sites in eag (a) and CG6004 (b). The editing sites are shown in separate tracks and the corresponding codon changes are shown next to each edited site (for example, YC in eag means Tyrosine is changed to Cysteine).


Scale
chrX:
20 kb
dm3
14,860,000
14,870,000
14,880,000
14,890,000
14,900,000
eag - RNA Editing Sites
KR
KR
YC
ND
IV
ND
RG
SG
RefSeq Genes
eag
hiw
FlyBase Protein-Coding Genes
eag
eag


Scale
chr3L:
2 kb
dm3
11,761,000
11,762,000
11,763,000
11,764,000
11,765,000
CG6004 - RNA Editing Sites
Silent
DG
Silent
Silent
Silent
Silent
KE
Silent
Silent
Silent
DG
Silent
Silent
Silent
Silent
Silent
Silent
Silent
Silent
RefSeq Genes
Muc68D
FlyBase Protein-Coding Genes
CG6004

these transcripts can be challenging but the RNA molecules can be studied directly using RNA sequencing

Combinatorial RNA editing

In *Drosophila*, the high-throughput RNA sequencing methods have identified over 900 RNA editing sites in ~600 genes, two thirds of which have the potential to alter amino acid codons [Graveley et al., 2011]. In the ultracomplex gene experiments, I found two of the sites present in *Ndae1* to undergo RNA editing but when I examined the data from Graveley *et al.,* [2011], I found examples of complex cases where multiple RNA editing sites are present in *eag* and *CG6004* (Figure 4.1) and are separated by longer distances. The long-reads generated by ONT can be applied to study combinatorial RNA editing present in these genes and at higher coverage, the abundance of combinatorial RNA editing sites can be studied.

Non-coding RNA characterization

In addition to studying exon connectivity in mRNAs and RN editing, the long reads generated by ONT can also be applied to characterize long non-coding RNAs. Currently, the direct RNA sequencing protocol requires the presence of poly(A)-tails in the 3' end of RNA but RNA molecules lacking poly(A)-tails can be studied in two ways. In the first method, A-tails can be added to the target RNA using *E.coli* poly(A)-polymerase and the resulting A-tailed RNA can be subjected to regular library protocol. In the second method, instead of adding poly(A)-tails, a common 3'-adapter sequence that is complementary to nanopore RT adapters can be used. Each of these methods require optimization because the number of A's added in the tails should be controlled to avoid sequencing A-repeats and similarly, in the second method, ligation of adapters need to be optimized to reduce artefacts resulting from self-ligated products.

Works Cited

1. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56-65 (2012).
2. Allseq.com. http://allseq.com/knowledge-bank/sequencing-platforms/solid/. Accessed 05/29/2018.
3. Amann, R.I., Ludwig, W., Schleifer, K.H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation Microbiol Rev. 59(1):143-69 (1995).
4. An, J.Y. National human genome projects: an update and an agenda. Epidemiol. Health 39: e2017045 (2017).
5. Ansorge, W., Sproat, B.S., Stegemann, J., Schwager, C. A non-radioactive automated method for DNA sequence determination. J Biochem Biophys Methods. 13(6): 315-323 (1986).
6. Arezi, B., Hogrefe H.H. Escherichia coli DNA polymerase III epsilon subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT-PCR procedures. Anal Biochem. 360(1):84-91 (2007).
7. Atkinson, M.R., Deutscher, M.P., Kornberg, A., Russell, A.F., Moffatt, J.G. Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. Biochemistry 8(12): 4897-4904 (1969).
8. Bannister, A.J., Kouzarides, T. Regulation of chromatin by histone modifications Cell Res 21(3):381-95 (2011).
9. Barnes, W.M. DNA sequencing by partial ribosubstitution. J. Mol. Biol. 119(1):83-99 (1978).
10. Berg, P., Fancher, H. & Chamberlin, M. Symposium on Informational Macromolecules, eds. Vogel, H., Bryson, V. & Lampen, J. 0. The synthesis of mixed polynucleotides contained ribo- and deoxyribonucleotides by purified preparations of DNA polymerase from Escherichia coli. Academic Press, New York & London 467-483(1963).
11. Berg, J.S., Agrawal, P.B., Bailey, D.B.Jr, Beggs, A.H., Brenner, S.E., Brower, A.M., Cakici, J.A., Ceyhan-Birsoy, O., Chan, K., Chen, F., et al. Newborn sequencing in genomic medicine and public health Pediatrics 139(2):2016-2252 (2017).

12. Berg, P., Fancher, H. & Chamberlin, M. The synthesis of mixed polynucleotides contained ribo- and deoxyribonucleotides by purified preparations of DNA polymerase from *Escherichia coli*. In Symposium on Informational Macromolecules, eds. Vogel, H., Bryson, V. & Lampen, J. O. Academic Press, New York & London, pp. 467-483 (1963).

13. Bolisetty, M.T., Rajadinakaran, G., Graveley, B.R. Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biology 16:204 (2014).

14. Briese, T., Paweska, J.T., McMullan, L.K., Hutchison, S.K., Street, C., et al. Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever–Associated Arenavirus from Southern Africa. PLoS Pathog 5(5): e1000455 (2009).

15. Brown, N.L. A primed-synthesis method for ribosubstitution of DNA at a single site. FEBS Lett. 93(1):10-15 (1978).

16. Brown, J.B., Boley, N., Eisman, R., May, G.E., Stiober, M.H., et al. Diversity and dynamics of the Drosophila transcriptome. Nature 512(7515):393-399 (2014).

17. Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C., Franklin, R.B. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience. 6(3):1-10 (2017).

18. Brown CG, Clarke J. Nanopore development at oxford nanopore. Nat. Biotechnol. 34(8):810-811 (2016).

19. Buermans, H.P., den Dunnen, J.T. Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta. 1842(10):1932-1941 (2014).

20. Carter, J.M., Hussain, S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. Wellcome Open Res.2:23 (2017).

21. Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608-611 (2015).

22. Chang, Z., Li, G., Zhang, Y., Ashby, C., Liu, D., Cramer, C.L., Huang, X. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 16:30 (2015).

23. Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., et al. The transcriptional diversity of 25 Drosophila cell lines. Genome Res. 21:301-314 (2011).

24. Chidgeavadze, Z.G., Beabealashvilli, R.S., Atrazhev, A.M., Kukhanova, M.K., Azhayev, A.V., Krayevsky, A.A. 2', 3'-Dideoxy-3' aminonucleoside 5' triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. Nucleic Acids Res. 12(3):1671-1686 (1984).

25. Cho, I., Blaser, M.J. The Human Microbiome: at the interface of health and disease. Nat. Rev. Genet. 13(4):260-270 (2012).

26. Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 14(6):671-83 (2013).

27. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. Landscape of transcription in human cells. Nature 489(7414):101-8 (2012).

28. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T,R. STAR: Ultrafast universal RNA-seq aligner Bioinformatics. 29(1):15-21 (2013).

29. Driscoll, C.B., Otten, T.G., Brown, N.M., Dreher, T.W. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Stand Genomic Sci. 12:9 (2017).

30. Drmanac, R, Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327(5961):78-81 (2010).

31. Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., Relman, D.A. Diversity of the human intestinal microbial flora. Science 308(5728):1635-1638 (2005).

32. Eichler, E.E., Clark, R.A., She, X. An assessment of the sequence gaps: Unfinished business in a finished human genome. Nat. Rev. Genet. 5(5):345-54 (2004).

33. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133-8 (2009).

34. Eisenstein, M. Startups use short-read data to expand long-read sequencing market. Nat. Biotechnol. 33(5):433-435 (2015).

35. Englund, P.T. The 3'-terminal nucleotide sequences of T7 DNA. J. Mol. Biol. 66(2):209-224 (1972).

36. Englund, P.T. Analysis of Nucleotide Sequences at 3' Termini of Duplex Deoxyribonucleic Acid with the Use of the T4 Deoxyribonucleic Acid Polymerase. J. Mol. Biol. 246(10): 3269-3276 (1971).

37. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473(7345):43-49 (2011).

38. Fagnani, M., Barash, Y., Ip, J.Y., Misquitta, C., Pan, Q., Saltzman, A.L., Shai, O., Lee, L., Rozenhek, A., Mohammad, N., et al. Functional coordination of alternative splicing in the mammalian central nervous system. Genome Biol. 8(6):R108 (2007).

39. Faria, N.R., Sabino, E.C., Nunes, M.R., Alcantara, L.C., Loman, N.J., Pybus, O.G. Mobile real-time surveillance of zika virus in brazil. Genome Med. 8(1):97 (2016).

40. Fededa, J.P., Petrillo, E., Gelfand, M.S., Neverov, A.D., Kadener,S., et al. A Polar Mechanism Coordinates Different Regions of Alternative Splicing within a Single Gene. Mol Cell. 19(3):393-404 (2005).

41. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods 7:461-465 (2010).

42. Frith, M.C., Hamada, M., Horton, P., Parameters for accurate genome alignment. BMC Bioinformatics 11:80 (2010).

43. Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15(3):201-6 (2018).

44. Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8(6):469-77 (2011).

45. GenomeWeb https://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business#.WtNir4jwaUk. Accessed on 04/14/2018.

46. Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. Integrative analysis of the caenorhabditis elegans genome by the modENCODE project. Science 330(6012):1775-1787 (2010).

47. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. Genome sequencing identifies major causes of severe intellectual disability. Nature 511(7509):344-347 (2014).

48. Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh,P.J., et al. Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312(5778): 1355–1359 (2006).

49. Glauser, D.A., Johnson, B.E., Aldrich, R.W., Goodman, M.B. Intragenic alternative splicing coordination is essential for caenorhabditis elegans slo-1 gene function. Proc. Natl. Acad. Sci. U S A. 108(51):20790-20795 (2011).

50. Goodwin, S., McPherson, J.D., McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17(6):333-351 (2016).

51. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29(7):644-652 (2011).

52. Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., et al. The Developmental Transcriptome of Drosophila melanogaster. Nature 471(7339):473-479 (2011).

53. Graveley, B.R. Mutually exclusive splicing of the insect dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell 123(1):65-73 (2005).

54. Harel, T., Lupski, J.R. Genomic disorders 20 years on-mechanisms for clinical manifestations. Clin. Genet. 93(3):439-449 (2018).

55. Hatem, A., Bozdag, D., Toland, A.E., Catalyurek, U.V. Benchmarking short sequence mapping tools. BMC Bioinformatics 14:184 (2013).

56. Hayden, E.C. The $1,000 genome. Nature 507, 294-295 (2014).

57. Heather, J.M., Chain, B. The sequence of sequencers: The history of sequencing DNA. Genomics 107(1):1-8 (2016).

58. Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R., Zamir, A. Structure of a ribonucleic acid. Science 147(3664):1462-1465 (1965).

59. Huang, Y.F., Chen, S.C., Chiang, Y.S., Chen, T.H., Chiu, K,P. Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst. Biol. 6(2):S10 (2012).

60. Huber, H.E., McCoy, J.M., Seehra, J.S., Richardson, C.C. Human immunodeficiency virus 1 reverse transcriptase. template binding, processivity, strand displacement synthesis, and template switching. J. Biol. Chem. 264(8):4669-4678 (1989).

61. Hunkapiller, T., Kaiser, R.J., Koop, B.F., Hood, L. Large-scale and automated DNA sequence determination. Science 254(5028):59-67 (1991).

62. Hunter, T., Francke, B. In vitro polyoma DNA synthesis: Inhibition by 1-beta-d-arabinofuranosyl CTP. J. Virol. 15(4):759-775 (1975).

63. Hutchison, C.A. DNA sequencing: Bench to bedside and beyond. Nucleic Acids Res. 35(18):6227-6237 (2007).

64. Illumina Handbook. DNA sequencing methods collection. https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/dna-sequencing-methods-review-web.pdf. Accessed on 04/19/2018.

65. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-945 (2004).

66. Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res. 4:1075 (2015).

67. Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallieres, M., Tapial, J., Raj, B., O'Hanlon, D., et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell 159(7):1511-1523 (2014).

68. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology 36:338-345 (2018).

69. Jain, M., Olsen, H.E., Paten, B., Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17:239 (2016).

70. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36(4):338-345 (2018).

71. Jiang, F., Ren, J., Chen, F., Zhou, Y., Xie, J., et al. Noninvasive Fetal Trisomy (NIFTY) test: an advanced noninvasive prenatal diagnosis methodology for fetal autosomal and sex chromosomal aneuploidies. BMC Medical Genomics 5:57 (2012).

72. Kasianowicz, J.J., Brandin, E., Branton, D., Deamer, D.W. Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U S A. 93(24):13770-3 (1996).

73. Khotskaya, Y.B., Mills, G.B., Shaw, K.R.M. Next-Generation Sequencing and Result Interpretation in Clinical Oncology: Challenges of Personalized Cancer Therapy. Annual Review of Medicine 68:113-125 (2017).

74. Kitts, P.A., Church, D.M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R.G., Tatusova, T., Xiang, C., Zherikov, A., et al. Assembly: A resource for assembled genomes. NCBI Nucleic Acids Res. 44(D1):D73-80 (2016).

75. Klarmann, G.J., Schauber, C.A., Preston, B.D. Template-directed pausing of DNA synthesis by HIV-1 reverse transcriptase during polymerization of HIV-1 sequences in vitro. J. Biol. Chem. 268(13):9793-9802 (1993).

76. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., Teichmann, S.A. The technology and biology of single-cell RNA sequencing. Mol. Cell. 58(4):610-20 (2015).

77. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420-426 (2007).

78. Koren, S., Schatz, S.C., Walenz, B.P., Martin, J., Howard, J.T., et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology 30:693–700 (2012).

79. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M. Canu: scalabale and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. 27(5):722-736 (2017).

80. Kotewicz, M.L., Sampson, C.M., D'Alessio, J.M., Gerard, G.F. Isolation of cloned moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. Nucleic Acids Res. 16(1):265-277 (1988).

81. Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., et al. Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol. 32(3):261-266 (2014).

82. Kurian, A.W., Hare, E.E., Mills, M.A., Kingham, K.E., McPherson, L., et al. Clinical Evaluation of a Multiple-Gene Sequencing Panel for Hereditary Cancer Risk Assessment. J Clin Oncol. 32-19:2001-2009 (2014).

83. Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., Burge, C.B. RNA bind-n-seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell. 54(5):887-900 (2014).

84. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921 (2001).

85. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25 (2009).

86. Laver, T., Harrison, J., O' Neill, P.A., Moore, K., Farbos, A., et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 3:1-8 (2015).

87. Lee, C., Roy, M. Analysis of alternative splicing with microarrays: Successes and challenges. Genome Biol. 5(7):231,2004-5-7-231 (2004).

88. Levy, S.E., Myers, R.M. Advancements in next-generation sequencing. Annu. Rev. Genomics Hum. Genet. 17:95-115 (2016).

89. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 (2013).

90. Li, H. Minimap2: pairwise alignment for nucleotide sequences. arXiv:1708.01492 (2018).

91. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079 (2009).

92. Li, H., Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 11(5):473-83 (2010).

93. Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. The sequence and de novo assembly of the giant panda genome. Nature 463(7279):311-317 (2010).

94. Lipkin, W.I. The changing face of pathogen discovery and surveillance. Nat. Rev. Microbiol.11(2):133-41 (2013).

95. Liu, L., Li, Y., Li, S., Hu, N., He, Y., et al. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology vol. 2012, Article ID 251364, (2012).

96. Loman, N.J., Quinlan, A.R. Poretools: A toolkit for analyzing nanopore sequence data. Bioinformatics 30(23):3399-3401 (2014).

97. Mardis, E.R. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9:387-402 (2008).

98. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376-380 (2005).

99. Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., et al. The genome sequence of the SARS-associated coronavirus. Science 300(5624):1399-1404 (2003).

100. Maxam, A.M., Gilbert. W. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U S A. 74(2):560-4 (1977).

101. McCoy, R. C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., et al. Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PLoS One 9(9): e106689 (2014).

102. McKernan, K., Blanchard, A., Kotler, L., Costa, G. In Reagents, methods, and libraries for bead-based sequencing, ed McKernan K., et al. (Agencourt Bioscience Corp. Beverly, MA). (2006).

103. McManus, C.J., Duff, M.O., Eipper-Mains, J., Graveley, B.R. Global analysis of trans-splicing in drosophila. Proc. Natl. Acad. Sci. U S A. 107(29):12975-12979 (2010).

104. McManus, C.J., Graveley, B.R. RNA structure and the mechanisms of alternative splicing. Curr. Opin. Genet. Dev. 21(4):373-9 (2011).

105. Merker, J.D., Wenger, A.M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K.S., et al. Long-read genome sequencing identifies causal structural variation in a mendelian disease. Genet. Med. 20(1):159-63 (2018).

106. Merriman, B., Ion, Torrent R&D Team., Rothberg, J.M. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 33(23):3397-417 (2012).

107. Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., Polioudakis, D., Iyer, V.R., Hunicke-Smith, S., Swamy, S., et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. Rna 19(7):958-70 (2013).

108. Morgan, J.E., Carr, I.M., Sheridan, E., Chu, C.E., Hayward, B., Camm, N., Lindsay, H.A., Mattocks, C.J., Markham, A.F., Bonthron, D.T., et al. Genetic diagnosis of familial breast cancer using clonal sequencing. Hum. Mutat. 31(4):484-91 (2010).

109. Neves, G., Zucker, J., Daly, M., Chess, A. Stochastic yet biased expression of multiple dscam splice variants by individual cells. Nat. Genet. 36(3):240-6 (2004).

110. NHGRI, https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/. Accessed on 05/29/2018

111. Nilsen, W. T., Graveley, B.R. Expansion of the eukaryotic proteome by alternative splicing. Nature 463:457-463 (2010).

112. Oikonomopoulos, S., Wang, Y.C., Djambazian, H., Badescu, D., Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci Rep. 6:31602 (2016).

113. Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulous, G.A., Papanikoloau, N., et al. Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. Bioinform Biol Insights 9: 75–88 (2015).

114. Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., et al. Direct RNA sequencing. Nature 461(7265):814-818 (2009).

115. Ozsolak, F., Milos, P.M. Transcriptome profiling using single-molecule direct RNA sequencing. Methods Mol. Biol. 733:51-61 (2011).

116. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 15(2):256-78 (2014).

117. Pages H, Aboyoun P, Gentleman R and DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.34.1.

118. Park. J.W., Graveley, B.R. Complex alternative splicing. Adv Exp Med Biol.623:50-63 (2007).

119. Pitukkijronnakorn, S., Promsonthi, P., Panburana, P., Udomsubpayakul, U., Chittacharoen, A. Fetal loss associated with second trimester amniocentesis. Arch. Gynecol. Obstet. 2011 284(4):793-797 (2011).

120. Plocik, A.M., Graveley, B.R. New insights from existing sequence data: Generating breakthroughs without a pipette. Mol. Cell. 49(4):605-17 (2013).

121. Porreca, G.J. Genome sequencing on nanoballs. Nat. Biotechnol. 28(1):43-4 (2010).

122. Potter, J., Zheng, W., Lee, J. thermal stability and cDNA synthesis capability of SuperScript™ III reverse transcriptase. Focus 25.1; 19-24 (2003).

123. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341 (2012).

124. Quick, J., Quinlan, A.R., Loman, N.J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. Gigascience 3:22 (2014).

125. Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biology 16:114 (2015).

126. Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., et al. Real-time, portable genome sequencing for ebola surveillance. Nature 530(7589):228-32 (2016).

127. Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., Sloan, W.T. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods 6(9):639-41 (2009).

128. Quinlan, A.R., Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-842 (2010).

129. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (2018).

130. Rhoads, A., Au, F. K. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13(5):278-289 (2015).

131. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., Cohen, J.C. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. 39(4):513-6 (2007).

132. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348-52 (2011).

133. Rothberg, J.M., Leamon, J.H. The development and impact of 454 sequencing. Nat. Biotechnol. 26(10):1117-24 (2008).

134. Roy, C.K., Olson, S., Graveley, B.R., Zamore, P.D., Moore, P.J. Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNA ligation. Nat. Methods 13:4 (2015).

135. Ruffalo, M., LaFramboise, T., Koyuturk, M.. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27(20):2790-2796 (2011).

136. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Nucleotide sequence of bacteriophage φX174 DNA. Nature 265(5596):687-695 (1977).

137. Sanger, F. Sequences, sequences, and sequences. Annu. Rev. Biochem. 57:1-28 (1988).

138. Sanger, F., Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94(3):441-448 (1975).

139. Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., Smith, M. The nucleotide sequence of bacteriophage phiX174. J. Mol. Biol. 125(2):225-46 (1978).

140. Sanger, F., Donelson, J.E., Coulson, A.R., Kossel, H., Fischer, D. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. Proc. Natl. Acad. Sci. U S A. 70(4):1209-13 (1973).

141. Schadt, E.E., Turner, S., Kasarskis, A. A window into third-generation sequencing. Hum. Mol. Genet. 19(R2):R227-240 (2010).

142. Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., Zipursky, S.L. Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101(6):671-84 (2000).

143. Sharon, D., Tilgner, H., Grubert, F., Snyder, M. A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. 31(11):1009-1014 (2013).

144. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., et al. DNA sequencing at 40: past, present and future. Nature 550(7676): 345-353 (2017).

145. Shendure, J., Porreca, G., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., Church, G.M. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741);1728-1732 (2005).

146. Smith, L.M., Fung, S., Hunkapiller, M.W., Hunkapiller, T.J., Hood, L.E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res. 13(7):2399-2412 (1985).

147. Soden, S.E., Saunders, C.J., Willig, L.K., Farrow, E.G., Smith, L.D., Petrikin, J.E., LePichon, J.B., Miller, N.A., Thiffault, I., Dinwiddie, D.L., et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci. Transl. Med. 6(265):265ra168 (2014).

148. Sovic, I., Sikic, M., Wilm, A., Fenlon, S.N., Chen, S., Nagarajan, N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat. Commun. 7:11307 (2016).

149. Springer M. Applied Biosystems: Celebrating 25 years of Advancing Science. American Laboratory. 38(11);4-8 (2006).

150. Sun, W., You, X., Gogol-Doring, A., He, H., Kise, Y., Sohn, M., Chen, T., Klebes, A., Schmucker, D., Chen, W. Ultra-deep profiling of alternatively spliced drosophila dscam isoforms by circularization-assisted multi-segment sequencing. Embo. J. 32(14):2029-2038 (2013).

151. Teng, M., Love, M.I., Davis, C.A., Djebali, S., Dobin, A. et al. A benchmark for RNA-seq quantification pipelines. Genome Biology 17:74 (2016).

152. Thermofisher. https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing/ion-s5-specifications.html. Accessed on 04/14/2018.

153. Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C.D., Rasmussen, M., Snyder, M.P. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat. Biotechnol. 33(7):736-42 (2015).

154. Topol, E.J., Frazer, K.A. The resequencing imperative. Nat. Genet. 39(4): 439-440 (2007).

155. Trapnell, C., Williams, B.A., Pertea, G., Mortzavi, A., Kwan, G., et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat. Biotechnol. 28(5):511-515 (2010).

156. Trapnell, C., Pachter, L., Salzberg, S.L. TopHat: Discovering splice junctions with RNA-seq. Bioinformatics 25(9):1105-1111 (2015).

157. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel. H., Salzberg, S.L., Rinn, J.L., Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat. Protoc. 7(3):562-578 (2012).

158. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L. Transcript assembly and quantification by RNA-seq

reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28(5):511-515 (2010).

159. Treutlein, B., Gokce, O., Quake, S.R., Sudhof, T.C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. PNAS 111(13) E1291-E1299 (2014).

160. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43 (2004).

161. Van de Sande, J.H., Loewen, P.C., Khorana, H.G. Studies on polynucleotides. 118. A further study of ribonucleotide incorporation into deoxyribonucleic acid chains by deoxyribonucleic acid polymerase I of Escherichia coli. J. Biol. Chem. 247(19):6140-6148 (1972).

162. van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C. Ten years of next-generation sequencing technology. Trends Genet. 30(9):418-426 (2014).

163. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods 13(6):508-514 (2016).

164. Vilfan.I.D., Tsai, Y., Clark, T.A., Wegener, J., Dai, Q., et al. Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. Journal of Nanobiotechnology 11:8 (2013).

165. Villalva, C., Touriol, C., Seurat, P., Trempat, P., Delsol, G., Brousset, P. Increased yield of PCR products by addition of T4 gene 32 protein to the SMART PCR cDNA synthesis system. Biotechniques 31(1):81-83, 86 (2001).

166. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed on 03/19/2018.

167. Wick, R.R., Judd, L.M., Holt, K.E. Comparison of Oxford Nanopore basecalling tools. https://github.com/rrwick/Basecalling-comparison. Accessed on 04/14/2018.

168. Wickham, H. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York (2009).

169. Willig, L.K., Petrikin, J.E., Smith, L.D., Saunders, C.J., Thiffault, I., et al. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. Lancet Respir Med. 3(5):377-387 (2015).

170. Wu, R., Kaiser, A.D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J. Mol. Biol. 35(3):523-537 (1968).

171. Wu. R., Taylor. E. Nucleotide sequence analysis of DNA: Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA. J. Mol. Biol. 57:491-511 (1971).

172. Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., Adams, M.D., Sun, S. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData. Min. 5(1):6 (2012).

173. Zhang, S., Bernstein, S.I. Spatially and temporally regulated expression of myosin heavy chain alternative exons during drosophila embryogenesis. Mech. Dev. 101(1-2):35-45 (2001).

174. Zhao, C., Liu, F., Pyle, A.M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. RNA 24(2):183-195 (2018).